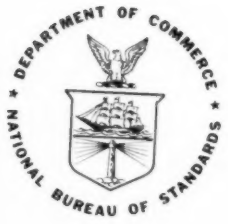
SECTION B



JOURNAL OF RESEARCH

of the NATIONAL BUREAU OF STANDARDS

JNBBAU 80B (2) 189-331, Apr.-June 1976

U.S. DEPARTMENT OF COMMERCE / National Bureau of Standards

Mathematical Sciences

Journal of Research

of the

National Bureau of Standards

B. Mathematical Sciences

Apr.–June • 1976

VOLUME • 80B

NUMBER • 2

Editor: Morris Newman

Associate Editor: F. W. Olver

Publication dates: Mar. 10, June 10, Sept. 10, Dec. 10, 1976

J. Res. Nat. Bur. Stand. (U.S.), B

JOURNAL OF RESEARCH

The National Bureau of Standards Journal of Research reports research and development in the fields of activity shown at right. Also included from time to time are survey articles on topics closely related to the Bureau's scientific and technical program.

The Journal is published in two separate sections as follows:

A. Physics and Chemistry

Contains papers of interest primarily to scientists working in these fields. Issued six times a year. Annual subscription: domestic, $17.00; foreign, $21.25. Single copy, varies.

B. Mathematical Sciences

Presents studies and compilations designed mainly for the mathematician and the theoretical physicist. Issued quarterly. Annual subscription: domestic, $9.00; foreign, $11.25. Single copy $2.25.

The National Bureau of Standards[1] was established by an act of Congress March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau consists of the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Institute for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of the Office of Measurement Services, the Office of Radiation Measurement and the following Center and divisions:

Applied Mathematics — Electricity — Mechanics — Heat — Optical Physics — Center for Radiation Research: Nuclear Sciences; Applied Radiation — Laboratory Astrophysics[2] — Cryogenics[2] — Electromagnetics[2] — Time and Frequency[2].

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement, standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; and develops, produces, and distributes standard reference materials. The Institute consists of the Office of Standard Reference Materials, the Office of Air and Water Measurement, and the following divisions:

Analytical Chemistry — Polymers — Metallurgy — Inorganic Materials — Reactor Radiation — Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations leading to the development of technological standards (including mandatory safety standards), codes and methods of test; and provides technical advice and services to Government agencies upon request. The Institute consists of the following divisions and Centers:

Standards Application and Analysis — Electronic Technology — Center for Consumer Product Technology: Product Systems Analysis; Product Engineering — Center for Building Technology: Structures, Materials, and Life Safety; Building Environment; Technical Evaluation and Application — Center for Fire Research: Fire Science; Fire Safety Engineering.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in improving cost effectiveness in the conduct of their programs through the selection, acquisition, and effective utilization of automatic data processing equipment; and serves as the principal focus within the executive branch for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Institute consists of the following divisions:

Computer Services — Systems and Software — Computer Systems Engineering — Information Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal Government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System; provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data — Office of Information Activities — Office of Technical Publications — Library — Office of International Relations — Office of International Standards.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.

[2] Located at Boulder, Colorado 80302.

U.S. DEPARTMENT OF COMMERCE
Elliot L. Richardson, Secretary

NATIONAL BUREAU OF STANDARDS
Ernest Ambler, Acting Director

Order all publications from the Superintendent of Documents
U.S. Government Printing Office, Washington, D.C. 20402.

The Secretary of Commerce has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this Department. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1976.

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

The Dependence of Inspection-System Performance on Levels of Penalties and Inspection Resources

A. J. Goldman and M. H. Pearl

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(April 8, 1976)

This paper presents three simple mathematical models, all of game-theoretic type, dealing with an inspector-inspectee relationship. The inspectee always tries to maximize his net gain, which is the amount he obtains by "cheating" less the amount he is penalized when caught. The first model assumes a zero-sum payoff and so the inspector tries to minimize the inspectee's net gain. In the second model, the inspector tries to deter cheating without concern for the extraction of penalties. In the third model we assume that the probabilistic pattern of the inspector's strategy is known to the inspectee and that the inspector constructs his strategy with this in mind. Each of these models is analyzed and optimal solutions are obtained. Several simple examples are presented to show the relation between the level of cheating and the levels of inspection resources and penalty.

Key words: Inspection; mathematical model; regulation; strategy; theory of games.

1. Introduction

This paper contains the formulation and analysis of three simple mathematical models, of game-theoretic type, oriented toward certain important aspects of the inspector-inspectee relationship. These models were initially designed in the course of a study performed for the NBS Office of Weights and Measures, while seeking techniques that might be useful to state "W&M" agencies in employing their resources to achieve best protection of consumers against loss due to malfunctioning weighing and measuring devices used in retail trade. However, they also appear relevant to many other situations involving an inspector-inspectee relationship in which it is meaningful to assign numerical values to the detection and to the nondetection of malfunction.

For example, food packaging plants (the inspectees) are required to assure through self-policing that no more than a specified level of foreign matter (in the form either of less expensive substitutes or of "dirt") enters into the food they package. An agency of government (the inspector) is charged with guaranteeing to the public that these levels are not exceeded. If we can assign numerical values to (1) the benefit accruing to the packager (for example, in reduced operating expenses) for permitting systematic violation of the maximal level of foreign matter, and, (2) the penalty for being found in violation, then the models and analysis of this paper might well be pertinent. Similar comments can be made with "food purity" replaced by "consumer product safety."

The type of model at which we aim may also prove applicable to situations not customarily described in terms of inspectors and inspectees. For example, a department store (the inspector) wishes to combat thievery among its workforce (the inspectees) and establishes an internal security section with a fixed amount of resources. If numerical values can be specified for successful and for unsuccessful stealing, then models like those of this paper might be used to minimize the department store's loss. Many other scenarios can be formulated for which such models could describe "optimal strategies" for both the inspector and the inspectee.

Although the sort of analysis initiated below promises to have broad scope, it should be stated at the outset that the present models are not detailed and highly realistic ones, capable of giving "practical answers." Rather, they are intended both as preliminary explorations of how certain

AMS Subject Classification: 90D40.

issues might be framed for quantitative analysis, and as concrete targets for criticisms and suggestions leading to superior versions.

The models ignore many features of the inspection process, for the sake of focusing sharply and clearly on certain other features. Some of the ignored aspects will be explicitly identified later, but for general perspective it should be noted that our emphasis here is one-sidedly on the "consumer protection" or "deterrence of cheating" element in inspection operations, as distinguished from the "help the inspectee (merchant) keep his device accurate" theme.

The features of the situation on which the models are intended to focus are the following:

(a) A commercial establishment, containing one or more measuring devices which "meters" its transactions with customers, has an economic incentive to "cheat" in connection with that device. The ugly word "cheat" is used as a pithy abbreviation for any of (i) deliberately causing the device to malfunction in the direction economically advantageous to the establishment (for example, "short weighing"), or (ii) knowingly permitting such a malfunction (originating through natural causes) to go uncorrected, or (iii) employing such a malfunctioning device, which though not explicitly recognized as such, was not properly checked for malfunction. Different establishments, with different dollar-flows of transactions, will experience economic incentives of different magnitudes. The models contain a set of parameters (V_i) representing these magnitudes.

(b) The inspection agency typically has only a limited quantity of "inspection resources" at its disposal, a quantity likely to be too small for frequent coverage of all the devices in its jurisdiction. The models contain parameters, m and n, representing respectively, the inspection resources and the number of devices over which they must be spread.

(c) The detection of cheating leads to imposition of a "penalty" on the malefactor. This penalty might be the sum of (i) a monetary fine, (ii) the dollar-equivalent of interruption of the normal flow of business, and (iii) the economic loss due to diminished patronage following public disclosure of the cheating. The models include a parameter (P) representing the size of this penalty.

(d) Our purpose in this paper is to begin the development of additional methodological tools for designing more effective and efficient responses by society to the existence of cheating. Those responses include both a punitive element (whose intensity is measured by P) and a "policing" element (whose intensity is represented by m). By determining how the level of illicit activity depends on P and m, we hope to contribute to better understanding of the *effectiveness* of the two-element response. By exhibiting the tradeoffs between P and m in the above dependence, we hope to contribute to better understanding of the *efficiency* with which that response is allocated between the two elements, though full achievement of that goal will also require models of the "cost" (as well as the "performance output") associated with various combinations of P and m. These particular objectives, which of course remain relevant in a broad context of regulatory and general criminal-justice activity, have guided the development of the models to be described.

The mathematical formulation of a first model is given in the next section (sec. 2), and is accompanied by a discussion of some of that model's shortcomings in order to indicate directions for further research. The results from analyzing this Model 1 are presented in section 3, along with two simple illustrations; the analysis itself appears in section 4 (which the reader may prefer to omit). This model makes the "zero-sum" assumption—found in most game-theoretic literature—that the inspection agency's objectives are antithetical to those of the "inspectee," which implies in particular that the latter's loss through penalty-payment is the former's gain.

Section 5, in contrast, takes up a second model which is of nonzero-sum type; it treats the basic aim of the inspection activity as *deterrence* of cheating, with no separate value for the extraction of penalties after its detection. The "solution" of this Model 2 is presented and illustrated, with the supporting analysis deferred to section 6.

In section 7 we drop the usual game-theoretic assumption that the players select their strategies simultaneously (or, at least, that the strategy chosen by each player does not depend on the strategy chosen by his opponent). Since the inspection procedure is an ongoing one, it would be expected that the inspectee, sooner or later, would be able to discern the probabilistic pattern of the

inspector's choices. Thus, in Model 3 we assume that when the inspectee formulates his strategy he knows what the inspector's strategy will be. Moreover the inspector is aware of this and so constructs his strategy with it in mind. The solution of Model 3 was suggested by NBS colleague L.S. Joel; the results for Model 3 are presented in section 7 while section 8 supplies some of the longer derivations justifying these results.

2. Formulation of Model 1

This mathematical model takes the form of a 2-player zero-sum game. The "players" are the *inspector* (an aggregate representing the inspection agency) and the *inspectee* (an aggregate representing the establishments in which the measuring devices are used). There is some loss in reality through regarding these establishments as forming a single player with a single interest, but for the present model this potential turns out not to be realized.

The inspectee can either cheat, or not, for each of a set of devices $D_1, D_2, \ldots D_n$. The inspector selects devices for inspection, up to the limit of his resources. Detection of a cheat, if the device is inspected, is assumed certain. The data of the model are:

n = number of devices,
V_i = payoff to inspectee from cheating on D_i,
P = penalty against inspectee for each detection of cheating,
m = number of devices the inspector can examine ($m < n$).

A *strategy* for the inspectee is an n-component vector

$$\mathbf{c} = (c_1, c_2, \ldots, c_n)$$

in which

$c_i = 1$ if there is cheating on D_i,
$c_i = 0$ if there is no cheating on D_i.

The consideration of inspectee *mixed strategies* (probabilistic mixtures of ordinary strategies) is not necessary for the present model, i.e., the model has solutions which do not involve them, but we will also present those solutions which do include mixed inspectee strategies.

A strategy for the *inspector* is a specification of a subset M of the set $N = \{1,2, \ldots, n\}$ such that M has exactly m members (notation: $|M| = m$). The interpretation is that $i \in M$ if and only if D_i is inspected. It turns out that consideration of mixed strategies for the inspector is required in order to solve the present model. Initially we define such a mixed strategy to be a vector $\mathbf{p}$, with a component $p(M)$ for each subset M of N such that $|M| = m$, and the properties that

$$p(M) \geq 0, \ \sum_M p(M) = 1. \tag{2.1}$$

The interpretation is given by

$$p(M) = \text{Prob}\,[\{D_i : \ i \in M\} \text{ are the devices inspected}].$$

With each such $\mathbf{p}$ can be associated the quantities

$$\begin{aligned} p_i &= \text{Prob}\,[D_i \text{ is inspected}] \\ &= \sum \{p(M) : M \text{ contains } i\}. \end{aligned} \tag{2.2}$$

Clearly each p_i satisfies $0 \leq p_i \leq 1$. As will be proved in section 4, the relation $\sum_1^n p_i = m$ holds.

Conversely, consider any n-component vector $(p_1,p_2, \ldots, p_n)$ with

$$0 \leq p_i \leq 1, \ \sum_1^n p_i = m. \tag{2.3}$$

As will also be proved in section 4, there is then a mixed strategy $\mathbf{p}$ whose components $p(M)$ satisfy

$$p_i = \sum \{p(M) : M \text{ contains } i\}$$

for all $i \in N$. Thus we can and will regard a mixed strategy as a vector $\mathbf{p} = (p_1,p_2, \ldots, p_n)$ whose components p_i, interpreted as in (2.2), satisfy (2.3).

The net expected payoff to the inspectee, from device D_i, is the expected payoff from cheating minus the expected penalty, i.e.

$$V_i c_i - P(c_i p_i) = [V_i - P p_i]c_i.$$

Thus the total net expected payoff to the inspectee, if the two players choose respective mixed strategies $\mathbf{c}$ and $\mathbf{p}$, is

$$F(\mathbf{c},\mathbf{p}) = \sum_1^n (V_i - P p_i)c_i. \tag{2.4}$$

From the "zero-sum" assumption that the interests of the two players are diametrically opposed, it follows that $-F(\mathbf{c},\mathbf{p})$ is the expected payoff to the inspector.

Before proceeding (in the next section) to present the "solution" given by this model—i.e., the "optimal strategies" for both players—we note some of the model's limitations and imperfections:

(a) The zero-sum assumption of diametrically opposed interests is not quite right unless one thinks of society (whose agent is the inspector) as seeking vengeance rather than deterrence. A second game-theoretic model, intended to give better expression to the "deterrence" theme, is formulated and analyzed in section 5.

(b) The "cheat or no-cheat" dichotomy is clearly a severe idealization, ignoring as it does the possibility of introducing different degrees of bias into the devices (different amounts of cheating). To consider this dimension of the situation would also require formulating some mathematical representation of how the probability of detection depends on the degree of cheating, and could raise interesting questions concerning the role of the establishments' customers in performing part of the detection function (that of recognizing gross cheating). For the present we decline to explore these lines of generalization.

(c) Even with the above restriction, it seems odd to assume that cheating will be detected with certainty if the offending device is but inspected. This for example seems to rule out use of the model to examine the relative merits of training, or selecting, or otherwise encouraging inspectors to work more rapidly (in effect, increasing m), versus stressing the thoroughness or quality of their work (in effect, increasing he probability of cheat-detection). Fortunately, this limitation is only formal; a detection probability δ can be represented in the model simply by replacing the fixed penalty P, in (2.4), by the average penalty $P' = P\delta$.

(d) A quite natural extension of the model would be the replacement of P, in (2.4), by device-specific penalties P_i. (For example, the presence of site-dependent detection probabilities would lead, as above, to the use of $P\delta_i$ in place of P. Also, interruption of business or loss of patronage could be more serious at sites with a greater volume of transactions, suggesting the association of larger P_i's with larger V_i's.) This would raise no real problem were we content with solving the model numerically, but it does interfere with achieving the kind of nearly closed-form solution presented in the next section, and it definitely contradicts the desire to maintain in the model a simple clear-cut scalar quantity representing the intensity of society's sanction against cheating. So, this complication will be omitted from the present paper.

(e) If a serious level of cheating is detected, the inspector (government) may react with measures which are onerous to all inspectees (merchants), even those with properly functioning devices. Similarly, if detected cheating is publicized, honest as well as dishonest establishments may suffer from loss of public confidence. These considerations suggest an inadequacy in the way (2.4) associates penalty specifically to those devices at which cheating is discovered. Perhaps there should be a penalty P^* which is "activated" if cheating is discovered at *any* device (or at more than some "threshold" number of devices). This possibility is also left for future investigation; its inclusion would preclude expressing the payoff function in terms of the n p_i's rather than the much more numerous $p(M)$'s.

(f) The inspector must decide which subset of the n devices he will inspect, and the family of subsets which represent "allowable" outcomes of this decision are limited by the amount of inspection resources available. In the present model, that limitation is expressed by specifying the number (m) of inspections which can be undertaken. This is clearly an idealization of the real-world situation in which some inspections may (predictably) require more time than others, some inspection sites are remote from the majority so that visiting them substantially reduces the number of other devices which can be inspected during a given time period, etc. One would like to improve the model by incorporating a more realistic representation of the family of "allowable" subsets from which the inspector can choose.

The preceding list suggests a number of directions for further analysis. But Model 1 and the others in this paper, despite their evident deficiencies, are felt to represent a suitable first step in focusing on the issues of interest.

3. Results for Model 1

In describing the results for the model formulated above, it is convenient to maintain the notation $N = \{1,2, \ldots, n\}$, and also to employ the notations

$$T = \{i : V_i > P\},\ \bar{T} = N - T = \{i:\ V_i \leq P\}. \tag{3.1}$$

The letter "T" was chosen because the devices $\{D_i:\ i \in T\}$ are especially *t*empting choices for cheating.

The solution takes different forms according as

$$m < |T| + \sum \{V_i/P : i \in \bar{T}\} \qquad \text{(Case I)} \tag{3.2}$$

or its opposite

$$m \geq |T| + \sum \{V_i/P : i \in \bar{T}\} \qquad \text{(Case II)} \tag{3.3}$$

holds. If the V_i are thought of as fixed, then this division into cases can be construed as partitioning the first quadrant of (P,m)-space into two regions. The Case I region corresponds (as will be seen) to situations where the inspectee's optimal strategy is to cheat on *all* devices, and so this region may be interpreted as representing grossly inadequate societal responses to the threat of cheating.

For each of the two cases, we will present a strategy $\mathbf{c}^\circ$ for the inspectee, a mixed strategy $\mathbf{p}^\circ$ for the inspector, and a number F°, such that

$$F(\mathbf{c}^\circ,\mathbf{p}) \geq F^\circ \qquad \text{for all } \mathbf{p}, \tag{3.4}$$

$$F(\mathbf{c},\mathbf{p}^\circ) \leq F^\circ \qquad \text{for all } \mathbf{c}. \tag{3.5}$$

These relations of course imply that

$$F(\mathbf{c}^0,\mathbf{p}^0) = F^0. \tag{3.6}$$

Inequality (3.4) says that if the inspectee chooses strategy $\mathbf{c}^0$ then his payoff will be *at least* F^0 no matter what strategy the inspector picks, while (3.5) says that if the inspector chooses strategy $\mathbf{p}^0$ then the inspectee's payoff (the negative of his own) will be *at most* F^0 no matter what strategy the inspectee picks. On this basis, $\mathbf{c}^0$ and $\mathbf{p}^0$ are *optimal strategies* for the two players, and F^0 is the *value* of the game (to the inspectee); we take $\mathbf{F}^0 = \mathbf{F}(\mathbf{c}^0,\mathbf{p}^0)$ as the measure of average illicit net gains "predicted" by the model. This is the customary solution concept for zero-sum two-player games.

[A technical note to avoid a possible source of confusion: it can be proven from (3.4) and (3.5) that

$$F^0 = \max_{\mathbf{c}}\min_{\mathbf{p}} F(\mathbf{c},\mathbf{p}) = \min_{\mathbf{p}}\max_{\mathbf{c}} F(\mathbf{c},\mathbf{p}).$$

Thus the value can be defined without reference to any particular $\mathbf{c}^0$ and $\mathbf{p}^0$. With this done, condition (3.4) defines the notion of an optimal strategy $\mathbf{c}^0$ without reference to any $\mathbf{p}^0$, while (3.5) defines the notion of an optimal strategy $\mathbf{p}^0$ without reference to any $\mathbf{c}^0$. Such definitions are in general possible *only* for zero-sum two-player games.]

For Case I, as already stated, the optimal strategy for the inspectee is to cheat on all devices; formally, $\mathbf{c}^0$ is given by

$$c_i^0 = 1 \qquad \text{for all } i. \tag{3.7}$$

Any mixed strategy $\mathbf{p}^0$ for the inspector is optimal if it satisfies the "no overkill" condition

$$Pp_i^0 \leq V_i \qquad \text{for all } i \in \bar{T}. \tag{3.8}$$

The value is given by

$$F^0 = \sum_1^n V_i - Pm. \tag{3.9}$$

In Case II, it is convenient to partition $\bar{T}$ as $E \cup U$, where

$$E = \{i:\ V_i = P\}, \qquad U = \{i:\ V_i < P\}; \tag{3.10}$$

the symbols were chosen as the first letters of "equal" and "untempting." If strict inequality holds in (3.3), then the optimal strategies for the inspectee involve cheating on the tempting devices, but not on the untempting ones, i.e.,

$$c_i = 1 \qquad \text{for } i \in T, \tag{3.11}$$

$$c_i = 0 \qquad \text{for } i \in U, \tag{3.12}$$

$$c_i \text{ unspecified} \qquad \text{for } i \in E. \tag{3.13}$$

The phrasing of (3.13) was chosen to be compatible with the admission of mixed strategies for the inspectee (i.e., "$0 \leq c_i \leq 1$" rather than "$c_i = 0$ or 1"). In the special case that equality holds in (3.3), there is an additional family of optimal strategies obtained by choosing a number c with $0 < c \leq 1$ and (with (3.11) retained) replacing (3.12-3.13) with

$$c_i = c \qquad \text{for } i \in U, \tag{3.12'}$$

$$c_i \geq c \qquad \text{for } i \in E. \tag{3.13'}$$

The optimal strategies for the inspector require always inspecting the tempting devices and avoiding "underkill" on the others, i.e.,

$$p_i^\circ = 1 \qquad \text{for } i \in T, \qquad Pp_i^\circ \geq V_i \qquad \text{for } i \in \bar{T} \tag{3.14}$$

which implies $p_i^\circ = 1$ for $i \in E$. The value (whether (3.12–3.13) or (3.12'–3.13') is used) is

$$F^\circ = \sum \{V_i : i \in T\} - P|T|. \tag{3.15}$$

EXAMPLE 1: *Equal-Sized Firms*. This title is a somewhat free translation of the scenario-defining assumption that all V_i have a common value V. Then (3.2) and (3.9) give

$$F^\circ = nV - Pm \qquad \text{if } m < |T| + (V/P)\{n - |T|\}, \tag{3.16}$$

while (3.3) and (3.15) give

$$F^\circ = (V - P)|T| \qquad \text{if } m \geq |T| + (V/P)\{n - |T|\}. \tag{3.17}$$

If $P < V$ then $|T| = n$, so that (since we assume $m < n$) the defining condition of (3.16) holds and hence (3.16) rather than (3.17) is the case. If $V \leq P$ then $|T| = 0$, so that either (3.16) or (3.17) holds depending on whether or not $P < nV/m$. Our main interest is not in the optimal strategies but rather in the dependence of the "net illicit gain" F°, a measure of (imperfect) performance by the inspection system, as a function of P and m. This dependence is summarized in figure 1, in which F° is continuous across the curve (an equilateral hyperbola) separating the two regions shown. The notation $V_{\text{tot}} = \sum_1^n V_i (= nV)$ is used.

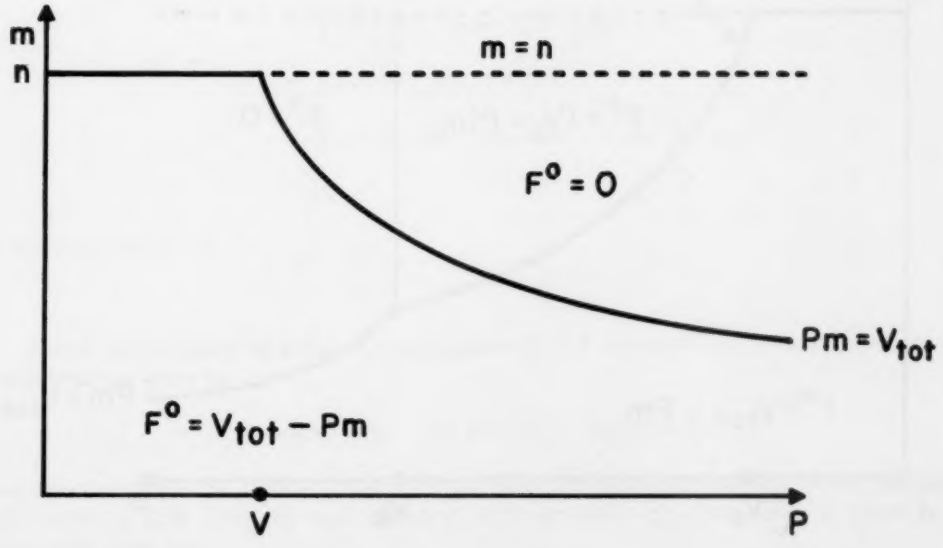


FIGURE 1. *Net illicit gain (equal-sized firms).*

EXAMPLE 2: *Big Firms, Small Firms*. Here there are n_b "big firms" each with $V_i = V_b$, and n_s "small firms" each with $V_i = V_s$, where $V_b > V_s$. We have $V_{tot} = n_bV_b + n_sV_s$. If $P < V_s$, so that $|T| = n$ and $\bar{T}$ is empty, then (3.2) and (3.9) give (since $m < n$)

$$F^o = V_{tot} - Pm.$$

If $V_s \le P < V_b$, so that $|T| = n_b$ and $|\bar{T}| = n_s$, then (3.2) and (3.9) give

$$F^o = V_{tot} - Pm \qquad \text{if } m < n_b + (V_s/P)n_s,$$

while (3.3) and (3.15) give

$$F^o = (V_b - P)n_b \qquad \text{if } m \ge n_b + (V_s/P)n_s. \tag{3.18}$$

If $P > V_b$, so that T is empty and $|\bar{T}| = n$, then (3.2) and (3.9) give

$$F^o = V_{tot} - Pm \qquad \text{if } m < V_{tot}/P,$$

while (3.3) and (3.15) give

$$F^o = 0 \qquad \text{if } m \ge V_{tot}/P.$$

These results are summarized in figure 2. F^o is continuous across each of the three boundary curves. Note that the region corresponding to (3.18) is "inspection-saturated"; a stiffer penalty, rather than more inspection activity, is needed to reduce the inspectee's expected payoff F^o from his optimal strategy (cheat at the big firms, but not at the small ones). (See table 1.)

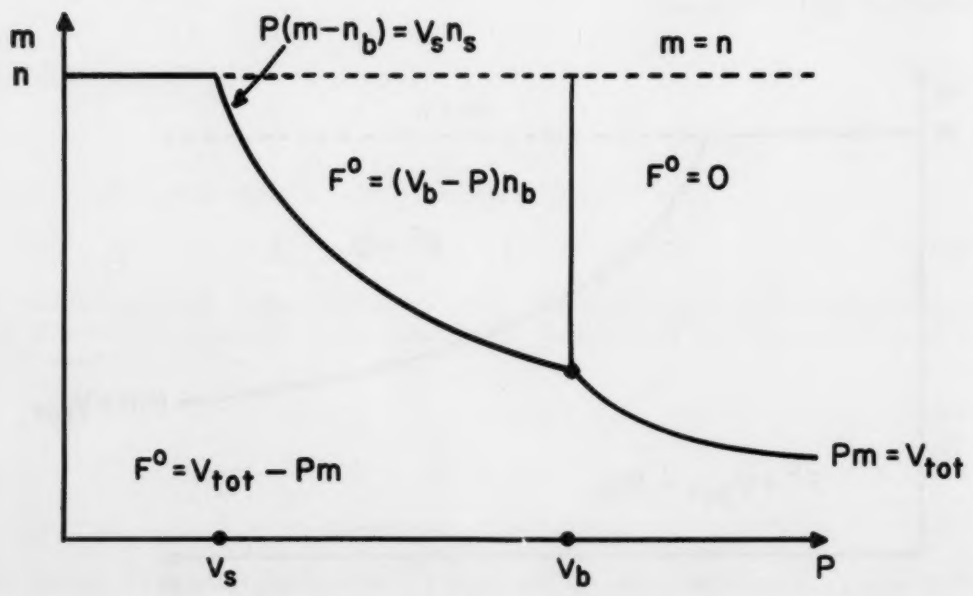


FIGURE 2. *Net illicit gain (big firms, small firms).*

TABLE 1. *Results for Model 1*

	Case Definition	Inspectee's Strategy	Inspector's Strategy	Payoff
Case I	$m < \lvert T\rvert + \sum\{V_i/P:\ i \epsilon \bar{T}\}$	$c_i^\dagger = 1$ for all i	$p_i^\dagger \le V_i/p$ for $i \epsilon \bar{T}$ $0 \le p_i^\dagger \le 1$ for $i \epsilon T$	$F^\circ = V_{\text{tot}} - Pm$
Case II	$m \ge \lvert T\rvert + \sum\{V_i/P:\ i \epsilon \bar{T}\}$	$c_i^\dagger = 1$ for $i \epsilon T$ $c_i^\dagger = 0$ for $i \epsilon U$ $0 \le c_i^\dagger \le 1$ for $i \epsilon E$	$p_i^\dagger = 1$ for i ge T $p_i^\dagger \ge V_i/P$ for $i \epsilon T$	$F^\circ = \sum\{V_i : i \epsilon T\}$ $- P\lvert T\rvert$

4. Analysis for Model 1

This section contains the mathematical justification of the results presented in section 3. Readers preferring to do so can go directly to section 5 without loss of continuity.

A first technical point to be settled is the permissability of passing from the original definition of a mixed strategy for the inspector, namely that of a vector with components $p(M)$ satisfying (2.1)—where M ranges over all subsets of N with $|M| = m$—to the subsequent definition as a vector $\mathbf{p}$ with components p_i satisfying (2.3). It will first be shown that each collection $\{p(M)\}$ obeying (2.1) leads via (2.2) to a set of p_i obeying (2.3); then it will be proved that each such $\mathbf{p} = (p_i)$ arises from such a collection $\{p(M)\}$.

For the first purpose, consider any collection $\{p(M)\}$ obeying (2.1), and let $a_i = 1$ for all $i \epsilon N$. With p_i defined by (2.2), we clearly have $p_i \ge 0$ as well as

$$p_i = \sum\{p(M):\quad M \text{ with } i \epsilon M\} \le \sum_M p(M) = 1,$$

and also

$$\begin{aligned}\sum_1^n p_i &= \sum_i a_i \sum\{p(M):\ M \text{ with } i \epsilon M\}\\ &= \sum\{a_i\, p(M):\ \text{pairs } (i,M) \text{ with } i \epsilon M\}\\ &= \sum_M p(M) \sum\{a_i:\ i \epsilon M\}\\ &= \sum_M p(M) \cdot m \qquad (\text{since } a_i = 1 \text{ and } |M| = m)\\ &= m,\end{aligned}$$

so that the p_i obey (2.3).

For the second purpose define, for each subset Q of N for which $|Q| = m$, a vector π_Q whose components are given by

$$\pi_Q(M) = 0 \qquad \text{for } M \neq Q; \qquad \pi_Q(Q) = 1.$$

Note that the π_Q are the vertices of the set (hypercube) S^* consisting of vectors π with components $p(M)$ satisfying (2.1). Next, for each subset Q of N for which $|Q| = m$, define a vector $\mathbf{p}_Q$ with components given by

$$p_{Qi} = 1 \qquad \text{if } i \epsilon Q; \; p_{Qi} = 0 \qquad \text{if } i \epsilon N - Q.$$

The set S, of all vectors $\mathbf{p}$ with components p_i satisfying (2.3), is the intersection of a hypercube with a hyperplane; its vertices must be the intersections of the hyperplane with the edges or vertices of the hypercube, from which it readily follows that the vertices of S are precisely the vectors $\mathbf{p}_Q$.

Finally, for each $i \in N$ and each subset M of N with $|M| = m$, let $a_{iM} = 1$ if $i \in M$ and $a_{iM} = 0$ otherwise. Then the relation (2.2) can be written

$$p_i = \sum_M a_{iM} p(M),$$

or, in linear transformation notation, as

$$\mathbf{p} = A\,\boldsymbol{\pi}.$$

It is readily vertified that $\mathbf{p}_Q = A\,\boldsymbol{\pi}_Q$. Now any vector $\mathbf{p} \in S$ can be written as a convex combination of the vertices of S, i.e.,

$$\mathbf{p} = \sum_Q \lambda_Q p_Q \qquad (\lambda_Q \geq 0,\ \sum_Q \lambda_Q = 1).$$

The vector

$$\boldsymbol{\pi} = \sum_Q \lambda_Q \boldsymbol{\pi}_Q,$$

as a convex combination of the vertices of S^*, is itself in S^*, and from the linearity of A we have

$$A\,\boldsymbol{\pi} = \sum_Q \lambda_Q A\,\boldsymbol{\pi}_Q = \sum_Q \lambda_Q \mathbf{p}_Q = \mathbf{p}.$$

This proves, as desired, that every $p \in S$ arises via (2.2) from some $\boldsymbol{\pi} \in S^*$.

To prove that the strategies presented in section 3 are indeed optimal, it is convenient to set

$$v_i = V_i/P \tag{4.1}$$

and to work, not with the payoff function F of (2.4), but rather with

$$f(\mathbf{c},\mathbf{p}) = F(\mathbf{c},\mathbf{p})/P = \sum_1^n (v_i - p_i)c_i. \tag{4.2}$$

Note that

$$T = \{i:\ \ v_i > 1\}, \qquad \bar{T} = \{i:\ \ v_i \leq 1\}. \tag{4.3}$$

We turn first to Case I, which according to (3.2) is characterized by

$$m < |T| + \sum \{v_i:\ \ i \in \bar{T}\}. \tag{4.4}$$

Using (3.9), we set

$$f^\circ = F^\circ/P = \sum_1^n v_i - m. \tag{4.5}$$

Following (3.7), let $\mathbf{c}^\circ$ be defined by $c_i^\circ = 1$ for all i. Then by (2.3) and (4.2)

$$f(\mathbf{c}^\circ,\mathbf{p}) = \sum_1^n (v_i - p_i) = \sum_1^n v_i - m = f^\circ$$

for all mixed strategies $\mathbf{p}$ for the inspector.

Next, consider any mixed strategy $\mathbf{p}^\circ$ such that $p_i^\circ \leq v_i$ for all $i \in T$. Then, since $\mathbf{p}^\circ \leq 1 < v_i$ for all $i \in T$, it follows that for all $\mathbf{c}$,

$$f(\mathbf{c},\mathbf{p}^\circ) = \sum_1^n (v_i - p_i^\circ)c_i$$

$$\leq \sum_1^n (v_i - p_i^\circ) = f^\circ. \tag{4.6}$$

Such mixed strategies $\mathbf{p}^\circ$ do exist, for if $m \leq \sum \{v_i\colon\ i \in \bar{T}\}$ they can be chosen with $p_i^\circ = 0$ for all $i \in T$, while if

$$\sum \{v_i\colon\ i \in \bar{T}\} < m < |T| + \sum \{v_i\colon\ i \in \bar{T}\},$$

then (4.4) assures the existence of some with $p_i^\circ = v_i$ for all $i \in \bar{T}$.

The preceding material shows that $(\mathbf{c}^\circ,\mathbf{p}^\circ,\mathbf{f}^\circ)$ give optimal strategies and the value for the game with payoff function f, so that $(\mathbf{c}^\circ,\mathbf{p}^\circ,F^\circ)$ give the corresponding information for the original game. Moreover, the indicated optimal strategies are the *only* ones. To see this, note first that the strictness of the inequality in (4.4) permits a choice of $\mathbf{p}^\circ$ in (4.6) for which $p_i^\circ < v_i$ for all $i \in \bar{T}$, and then equality will hold in (4.6) only for $\mathbf{c} = \mathbf{c}^\circ$. (A separate but easy argument is needed if $\bar{T}$ is empty.) Next, if $\mathbf{p}$ is any mixed strategy for which

$$I_1 = \{i\colon\ p_i > v_i, i \in \bar{T}\}$$

is nonempty, then define $\mathbf{c}$ by $c_i = 0$ for $i \in I_1$ and $c_i = 1$ for all $i \in N - I_1$. One obtains

$$\begin{aligned} f(\mathbf{c},\mathbf{p}) &= \sum \{v_i - p_i\colon\ i \in N - I_1\} \\ &= \sum_1^n (v_i - p_i) - \sum \{(v_i - p_i)\colon\ i \in I_1\} \\ &= f^\circ - \sum \{(v_i - p_i)\colon\ i \in I_1\} > f^\circ, \end{aligned}$$

and the existence of a $\mathbf{c}$ with this property rules out the optimality of $\mathbf{p}$.

Next we present the analysis for Case II, which according to (3.3) is characterized by

$$m \geq |T| + \sum \{v_i\colon\ i \in \bar{T}\} = |T \cup E| + \sum \{v_i\colon\ i \in U\}. \tag{4.7}$$

Using (3.12), we put

$$f^\circ = \sum \{v_i\colon\ i \in T\} - |T|. \tag{4.8}$$

With $\mathbf{c}^\circ$ defined by (3.11–3.13),

$$f(\mathbf{c}^\circ,\mathbf{p}) = \sum \{(v_i - p_i)\colon\ i \in T\} + \sum \{(1 - p_i)c_i\colon\ i \in E\}$$

$$\geq \sum \{(v_i - 1)\colon\ i \in T\} = f^\circ \tag{4.9}$$

for all $\mathbf{p}$. On the other hand, for $\mathbf{p}^\circ$ satisfying (3.14), i.e.,

$$p_i^\circ = 1 \qquad \text{for } i \in T;\ p_i^\circ \geq v_i \qquad \text{for } i \in \bar{T}, \tag{4.10}$$

we have

$$f(\mathbf{c},\mathbf{p}^\circ) \leq \sum \{(v_i - 1)\colon\ i \in T\} = f^\circ \tag{4.11}$$

for all $\mathbf{c}$. So $(\mathbf{c}^0,\mathbf{p}^0,f^0)$ form an optimal solution. The existence of mixed strategies $\mathbf{p}^0$ obeying (4.10) is assured by (4.7).

Once again, there are no other optimal inspector strategies. To see this, note first that equality holds in (4.9) only if

$$p_i = 1 \qquad \text{for all } i \in T,$$

which will therefore be assumed in the balance of this paragraph. If

$$I_2 = \{i: \ p_i < v_i, i \in \bar{T}\}$$

is nonempty then one can define $\mathbf{c}$ by

$$c_i = 1 \qquad \text{for } i \in T \cup I_2, \qquad c_i = 0 \qquad \text{for } i \in \bar{T} - I_2,$$

yielding

$$f(\mathbf{c},\mathbf{p}) = f^0 + \sum \{(v_i - p_i): \ i \in I_2\} > f^0$$

which shows that $\mathbf{p}$ is not optimal.

Next we inquire whether there are other optimal inspectee strategies; mixed strategies ($0 \leq c_i \leq 1$) are admitted in the discussion. Nite first that equality holds in (4.11) only if

$$c_i = 1 \qquad \text{for all } i \in T,$$

which is therefore assumed in the balance of this paragraph. Suppose that

$$K = \{k: \ c_k > 0, \qquad k \in U\}$$

is nonempty; we will attempt to deduce from this that $\mathbf{c}$ is not optimal. If strict inequality holds in (4.7), one can choose $\mathbf{p}$ to satisfy (4.10) with $p_k > v_k$ for all $k \in K$; thus

$$f(\mathbf{c},\mathbf{p}) = f^0 + \sum \{(v_k - p_k)c_k: \ k \in K\} < f^0,$$

so that $\mathbf{c}$ is not optimal. Now assume (4.7) holds with equality. If there exist distinct $k \in K$ and $i \in E \cup U$ with $c_i < c_k$, then choose any positive δ with

$$\delta < \min \{v_i, 1 - v_k\}$$

and define $\mathbf{p}$ by

$$p_j = 1 \qquad \text{for } j \in T \cup E - \{i\},$$

$$p_j = v_j \qquad \text{for } j \in U - \{i,k\},$$

$$p_i = v_i - \delta, \ p_k = v_k + \delta.$$

Then $\mathbf{p}$ is a mixed strategy, and

$$f(\mathbf{c},\mathbf{p}) = f^\circ + \delta c_i + (-\delta)c_k < f^\circ,$$

showing that $\mathbf{c}$ is not optimal. The only remaining possibility is that equality holds in (4.7) and, for some c with $0 < c \leq 1$,

$$c_i = c \qquad \text{for } i \in U;\ c_i \geq c \qquad \text{for } i \in E. \tag{4.12}$$

In this situation, for any $\mathbf{p}$,

$$\begin{aligned} f(\mathbf{c},\mathbf{p}) &= \sum \{(v_i - p_i):\ \ i \in T\} + \sum \{(1 - p_i)c_i:\ \ i \in E\} \\ &\quad + c\sum \{(v_i - p_i):\ \ i \in U\} \\ &= f^\circ + \sum \{(1 - p_i):i \in T\} + \sum \{(1 - p_i)c_i:\ \ i \in E\} \\ &\quad + c[\sum \{v_i:\ \ i \in U\} - \sum \{p_i:\ \ i \in U\}] \\ &\geq f^\circ + \sum \{(1 - p_i):\ \ i \in T\} + c\sum \{(1 - p_i):\ \ i \in E\} \\ &\quad + c[(m - |T \cup E|) - (m - \sum \{p_i:\ \ i \in T \cup E\})] \\ &= f^\circ + (1 - c) \sum \{(1 - p_i):\ \ i \in T\} \geq f^\circ, \end{aligned}$$

showing that $\mathbf{c}$ is indeed optimal.

One further technical question is whether the use of mixed strategies for the inspector is really necessary. Might not the inspector have "unmixed" (or "pure") strategies which are optimal? Such a strategy $\mathbf{p}^\circ$, besides obeying (2.3), would also satisfy

$$p_i^\circ = 0 \text{ or } 1 \qquad \text{for all } i \in N. \tag{4.12}$$

In Case II, we see from (4.10) that this would require all $p_i^\circ = 1$, so that (2.3) would be satisfied only in the uninteresting case $m = n$. For Case I, assuming no v_i exactly equal to 1, (4.12) and the condition $p_i^\circ \leq v_i$ for all $i \in \overline{T}$ imply that $p_i^\circ = 0$ for all $i \in \overline{T}$, so that (2.3) can be satisfied if and only if $|T| \geq m$, a condition which does not follow from the defining characteristic (4.4) of Case I. So mixed strategies are indeed needed in Case II, and also in Case I when $|T| < m$.

5. Formulation and Results for Model 2

In what follows, the payoff function for the inspectee is as before in (2.4), namely

$$F(\mathbf{c},\mathbf{p}) = \sum_1^n [V_i - Pp_i]c_i. \tag{5.1}$$

We now regard the inspector's aim, however, as that of minimizing the total loss (e.g., to consumers) as a result of cheating. The payoff function for the inspector is therefore taken initially to be

$$G(\mathbf{c},\mathbf{p}) = -\sum_1^n V_i c_i, \tag{5.2}$$

rather than the previous $-F(\mathbf{c},\mathbf{p})$.

The solution concept typically employed for such nonzero-sum (and "noncooperative") games is that of an *equilibrium point* (EP), namely a pair $(\mathbf{c}^0,\mathbf{p}^0)$ of strategies such that

$$F(\mathbf{c},\mathbf{p}^0) \le F(\mathbf{c}^0,\mathbf{p}^0) \qquad \text{for all } \mathbf{c}, \tag{5.3}$$

$$G(\mathbf{c}^0,\mathbf{p}) \le G(\mathbf{c}^0,\mathbf{p}^0) \qquad \text{for all } \mathbf{p}. \tag{5.4}$$

Such a pair has the "stability" property that if both players adopt them, then neither has incentive for a unilateral change in his strategy.

This approach is not helpful in the present situation, since *any* strategy $\mathbf{p}^0$ for the inspector is part of some EP, $(\mathbf{c}^0,\mathbf{p}^0)$. To show this, one need merely choose $\mathbf{c}^0$ to maximize $F(\mathbf{c},\mathbf{p}^0)$; then (5.3) is obeyed, and since $G(\mathbf{c},\mathbf{p})$ depends only on its first argument, (5.4) is also satisfied. Thus, the formulation provides no guidance for the inspector.

Two methods for escaping this unsatisfactory state of affairs have been considered. One of them, which involves abandonment of a strictly game-theoretic approach, is discussed in sections 7 and 8. The other ("Model 2"), to be treated in this section and the next one, introduces a slightly different objective for the inspector, namely to minimize the total loss due to *undetected* cheating. Thus (5.2) is replaced by

$$G(\mathbf{c},\mathbf{p}) = -\sum_1^n v_i c_i (1 - p_i). \tag{5.5}$$

A rationale for this objective—reasonable, but in the writers' opinion less than compelling—can be given in terms of a scenario in which society's response to cheating is so structured that the penalty for a *second* offense is prohibitive to the inspectee. (Thus, P now represents the penalty for a *first* (detected) offense.) For such a scenario, it is plausible that future cheating would occur precisely at the sites of current undetected cheating, so that its extent is measured by (5.5).

We now proceed to describe the equilibrium points $(\mathbf{c}^0,\mathbf{p}^0)$ of the game with payoff functions (5.1) and (5.5); the analysis supporting these results is given in section 6. The associated payoff values, $F^0 = F(\mathbf{c}^0,\mathbf{p}^0)$ and $G^0 = G(\mathbf{c}^0,\mathbf{p}^0)$, will also be presented.

The devices may be assumed numbered so that

$$V_1 \ge V_2 \ge \ldots \ge V_n. \tag{5.6}$$

The presence of ties among the V_i's can make this numbering somewhat arbitrary, and will lead to some complications (technical rather than substantive) in describing the solution of the model. In addition to the sets

$$T = \{i:\ V_i > P\}, \qquad \bar{T} = \{i:\ V_i \le P\}$$

defined earlier, it is convenient to define the sets

$$\begin{aligned} N_m &= \{1, 2, \ldots, m\}, & N_m^0 &= \{i:\ V_i = V_m\} \\ N_m^+ &= \{i:\ V_i > V_m\}, & N_m^- &= \{i:\ V_i < V_m\}. \end{aligned} \tag{5.7}$$

The three sets N_m^+, N_m^- and N_m^0 partition N; either or both of the first two might be empty, but N_m^0 contains at least the member m.

The discussion of Model 2, like that of Model 1, splits into cases. First assume

$$V_m > P \qquad \text{(Case I)}. \tag{5.8}$$

As in Case I of Model 1, all equilibrium points $(\mathbf{c}^\circ,\mathbf{p}^\circ)$ for this case involve cheating on all devices, i.e.,

$$c_i^\circ = 1 \qquad \text{for all } i. \tag{5.9}$$

The strategies $\mathbf{p}^\circ$ involve always inspecting those devices more tempting than the mth, and never inspecting those less tempting, i.e.,

$$p_i^\circ = 1 \qquad \text{for } i \in N_m^+, \qquad p_i^\circ = 0 \qquad \text{for } i \in N_m^-, \tag{5.10}$$

$$0 < p_i < 1 \qquad \text{for } i \in N_m^\circ, \ \sum \{p_i^\circ : i \in N_m^\circ\} = m - |N_m^+|. \tag{5.11}$$

Substituting (5.9—5.11) into (5.1) and (5.5), we find that the inspectee's payoff is

$$F^\circ = \sum_1^n V_i - Pm, \tag{5.12}$$

the same as for Case I of Model 1, and similarly the inspector's payoff is

$$G^\circ = -[V_m \{n - |N_m^-| - m\} + \sum \{V_i : \ i \in N_m^-\}]. \tag{5.13}$$

Next, assume

$$V_m = P \qquad \text{(Case II)}. \tag{5.14}$$

There are two subcases. If m is not the last member of N_m°, i.e.,

$$|N_m^+| + |N_m^\circ| > m, \qquad \text{(Case IIA)} \tag{5.15}$$

then the equilibrium points and payoff functions are given by the same formulas (5.9—5.13) as for Case I. But if

$$|N_m^+| + |N_m^\circ| = m, \qquad \text{(Case IIB)} \tag{5.16}$$

then the equilibrium points require always cheating on both the tempting and untempting devices, e.g.,

$$c_i^\circ = 1 \qquad \text{for } i \in N_m^+ = T, \tag{5.17}$$

$$c_i^\circ = 1 \qquad \text{for } i \in N_m^- = U, \tag{5.18}$$

together with a "floor" under the frequency of cheating on the other devices:

$$c_i^\circ \geq V_{m+1}/P \qquad \text{for } i \in N_m^\circ = E. \tag{5.19}$$

The inspection policies are given by

$$p_i^\circ = 1 \qquad \text{for } i \in N_m^+ \cup N_m^\circ = T \cup E, \tag{5.20}$$

$$p_i^\circ = 0 \qquad \text{for } i \in N_m^- = U. \tag{5.21}$$

The corresponding values of the payoff functions are

$$F^0 = \sum_1^n V_i - Pm, \tag{5.22}$$

as before, and

$$G^0 = -\sum \{V_i: \quad i \in N_m^-\}, \tag{5.23}$$

which in view of (5.16) "matches" with (5.13).

The remaining case is

$$V_m < P \qquad \text{(Case III)}. \tag{5.24}$$

If the condition

$$m \geq |T \cup E| + \sum \{V_i/P: \quad i \in U\} \qquad \text{(Case IIIA)} \tag{5.25}$$

holds, there is a family of equilibrium points given by

$$c_i^0 = 1 \qquad \text{for } i \in T, \tag{5.26}$$
$$0 \leq c_i^0 \leq 1 \qquad \text{for } i \in E, \tag{5.27}$$
$$c_i^0 = 0 \qquad \text{for } i \in U, \tag{5.28}$$
$$p_i^0 = 1 \qquad \text{for } i \in T \cup E \tag{5.29}$$
$$p_i^0 \geq V_i/P \qquad \text{for } i \in U. \tag{5.30}$$

The associated payoff values are

$$F^0 = \sum \{V_i: \quad i \in T\} - P|T|, \tag{5.31}$$

$$G^0 = 0. \tag{5.32}$$

A second and more complicated type of equilibrium point exists when

$$m \leq |T \cup E| + \sum \{V_i/P: \quad i \in U\}. \qquad \text{(Case IIIB)}. \tag{5.33}$$

For its description, let

$$L = \{i: \quad V_m < V_i < P\}, \tag{5.34}$$

so that N_m^+ is partitioned as $T \cup E \cup L$ and U as L $\cup$ $N_m^0 \cup N_m^-$. Furthermore, divide N_m^- into *blocks*; a "block" is defined to be a maximal succession of integers i with a common V_i-value. Let B_b denote the bth block, and V_b^* its associated V_i-value; the blocks are numbered B_1 through B_β in descending order of V_b^*. It will also be convenient to regard N_m^0 as an initial block B_0, so that $V_0^* = V_m$. Because $T \cup E \cup L$ does not contain m,

$$\Delta = m - |T \cup E| - \sum \{V_i/P: \quad i \in L\} > 0. \tag{5.35}$$

Choose the smallest integer α for which

$$\Delta \leq \sum_{b=0}^{\alpha} |B_b| V_b^*/P; \tag{5.36}$$

by (5.33), such an α must exist. Also, let

$$\lambda = \Delta - \sum_{b=0}^{\alpha-1} |B_b| V_b^*/P, \tag{5.37}$$

so that

$$0 < \lambda \leq |B_\alpha| V_\alpha^*/P.$$

Next, choose c_m° so that

$$c_m^\circ = V_\alpha^*/V_m \qquad \text{if } \lambda < |B_\alpha| V_\alpha^*/P, \tag{5.38}$$

$$V_{\alpha+1}^*/V_m < c_m^\circ \leq V_\alpha^*/V_m \qquad \text{if } \lambda = |B_\alpha| V_\alpha^*/P, \tag{5.39}$$

where in (5.39) we set $V_{\alpha+1}^* = 0$ if $\alpha = \beta$. Then an equilibrium point $(\mathbf{c}^\circ, \mathbf{p}^\circ)$ is given by

$$c_i^\circ = 1 \qquad \text{for } i \in T, \tag{5.40}$$

$$c_i^\circ \geq c_m^\circ V_m/P \qquad \text{for } i \in E, \tag{5.41}$$

$$c_i^\circ = c_m^\circ V_m/V_i \qquad \text{for } i \in L, \tag{5.42}$$

$$c_i^\circ = c_m^\circ V_m/V_i \qquad \text{for } \begin{cases} i \in B_b & \text{with } 0 \leq b < \alpha \\ i \in B_\alpha & \text{if } \lambda = |B_\alpha| V_\alpha^*/P, \end{cases} \tag{5.43}$$

$$c_i^\circ = 1 \qquad \text{for } \begin{cases} i \in B_\alpha & \text{if } \lambda < |B_\alpha| V_\alpha^*/P \\ i \in B_b & \text{with } \alpha < b \leq \beta. \end{cases} \tag{5.44}$$

$$p_i^\circ = 1 \qquad \text{for } i \in T \cup E, \tag{5.45}$$

$$p_i^\circ = V_i/P \qquad \text{for } i \in L, \tag{5.46}$$

$$p_i^\circ = V_i/P \qquad \text{for } \begin{cases} i \in B_b & \text{with } 0 \leq b < \alpha \\ i \in B_\alpha & \text{if } \lambda = |B_\alpha| V_\alpha^*/P, \end{cases} \tag{5.47}$$

$$p_i^\circ \leq V_i/P \qquad \text{for } i \in B_\alpha \quad \text{if } \lambda < |B_\alpha| V_\alpha^*/P, \tag{5.48}$$

$$p_i^\circ = 0 \qquad \text{for } i \in B_b \quad \text{with } \alpha < b \leq \beta, \tag{5.49}$$

$$\sum \{p_i^\circ : \ i \in B_\alpha\} = \lambda. \tag{5.50}$$

The associated payoffs are given by

$$F^\circ = \sum \{V_i : \ i \in T\} - P|T| + \sum_{b \geq \alpha} |B_b| V_b^* - P\lambda = V_{\text{tot}} - Pm, \tag{5.51}$$

$$G^\circ = -[\{\sum_{b>\alpha} |B_b| V_b^* + \{n - m - \sum_{b>\alpha} |B_b|\} c_m^\circ V_m\}]. \tag{5.52}$$

Note that in (5.52), the factor $c_m^\circ V_m$ in the last summand is equal to V_α^* if $\lambda < |B_\alpha| V_\alpha^*/P$. But if

$$\lambda = |B_\alpha| V_\alpha^*/P, \tag{5.53}$$

then $c_m^0 \mathbb{V}_m$ has as range of variation the interval $(V_{\alpha+1}^*, V_\alpha^*)$ determined by (5.39), so that G^0 as given by (5.52) has a corresponding range of variation. This illustrates the general theoretical possibility, in a nonzero sum game, for different equilibrium points (i.e., "solutions") to have different associated payoff levels. The present model exhibits this phenomenon only for the particular "coincidence" described by (5.53).

It should also be observed that Cases IIIA and IIIB may coexist, giving rise to different equilibrium points. This occurs precisely when

$$m = |T \cup E| + \sum\{V_i/P: \quad i \in U\}, \tag{5.54}$$

which in the context of Case IIIB implies that

$$\alpha = \beta, \qquad \lambda = |B_\beta|V_n/P, \tag{5.55}$$

so that the above-mentioned nonuniqueness of G^0 occurs.

The general results presented above will now be illustrated using the same two special scenarios employed for Model 1, namely the "equal-sized firms" case (all $V_i = V$) and the "big firms, small firms" scenario ($V_i = V_b$ for $1 \le i \le n_b$, $V_i = V_s$ for $n_b < i \le n_b + n_s = n$, $V_b > V_s$).

EXAMPLE 1: *Equal-Sized Firms*. Here $|N_m^+ \cup N_{\bar{m}}^-| = 0$ and $|N_m^0| = n$. Since $n > m$, comparison with (5.16) shows that Case IIB is ruled out. If $V \ge P$, so that Case I or Case IIA holds, then (5.12) and (5.13) give

$$F^0 = V_{\text{tot}} - Pm, \; G^0 = -V(n-m) \qquad (\text{if } V \ge P). \tag{5.56}$$

Now suppose Case III is in effect, i.e., $V < P$. Then $|T \cup E \cup L| = 0$. By reference to (5.25), (5.31) and (5.32), Case IIIA gives rise to

$$F^0 = G^0 = 0 \qquad (\text{if } V < P \text{ and } Pm \ge V_{\text{tot}}). \tag{5.57}$$

For Case IIIB, one has

$$\alpha = \beta = 0, \qquad \Delta = \lambda = m, \qquad |B_\alpha|V_\alpha^*/P = V_{\text{tot}}/P,$$

so that (5.51) and (5.52), together with (5.38) and (5.39), yield

$$F^0 = V_{\text{tot}} - Pm \qquad (\text{if } V < P \text{ and } Pm \le V_{\text{tot}}), \tag{5.58}$$

$$G^0 = -V(n-m) \qquad (\text{if } V < P \text{ and } Pm < V_{\text{tot}}), \tag{5.59}$$

$$-V(n-m) \le G^0 \le 0 \qquad (\text{if } V < P \text{ and } Pm = V_{\text{tot}}). \tag{5.60}$$

The dependence of G^0 on P and m is summarized in figure 3; the dependence of F^0 is given by the earlier figure 1. Note, in figure 3, that G^0 is *not* continuous across the hyperbola $Pm = V_{\text{tot}}$; as shown by (5.60), it is not single-valued on this curve, and its range of variation at each point of the curve is precisely the interval between its limits as the curve is approached from within each of the two regions identified in the figure. In the lower region, G^0 does not depend on P; this can be interpreted as a situation of "penalty saturation," in which stiffer penalties yield no improvement unless accompanied by greater inspection resources.

EXAMPLE 2: *Big Firms, Little Firms*. The situation here is more complicated. First suppose $V_b > V_s \ge P$. Then either Case I or Case II applies, so that by (5.12)

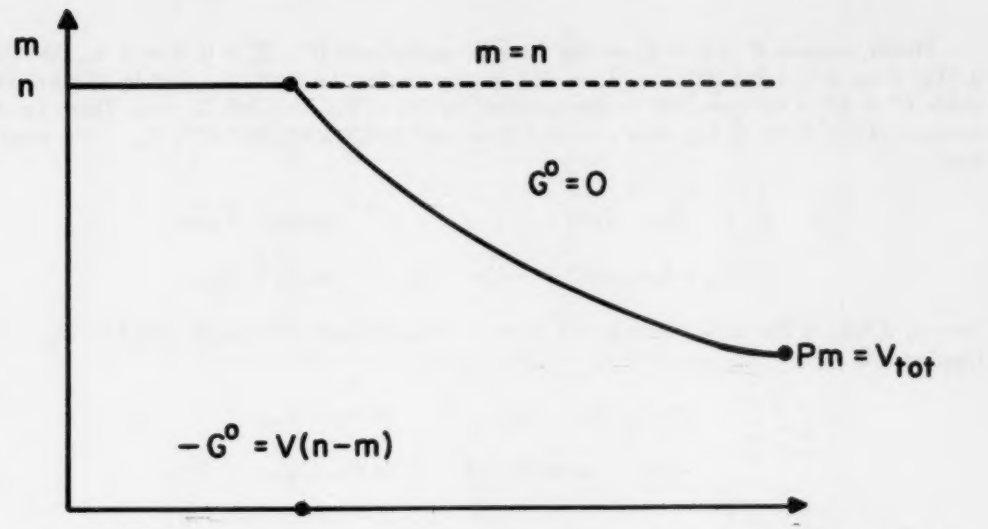


FIGURE 3. *Undetected illicit activity (equal-sized firms).*

$$F^\circ = V_{\text{tot}} - Pm \qquad (\text{if } V_b > V_s \geq P). \tag{5.61}$$

If $m \leq n_b$, then $V_m = V_b$, $|N_m^\circ| = n_b$ and $|N_m^-| = n_s$; Case I applies, and (5.13) yields

$$G^\circ = -(V_{\text{tot}} - V_b m) \qquad (\text{if } V_b > V_s \geq P \text{ and } m \leq n_b). \tag{5.62}$$

If $m > n_b$, then $V_m = V_s$, $|N_m^+| = n_b$, and $|N_m^\circ| = n_s$; Case IIB is ruled out, and (5.13) yields

$$G^\circ = -V_s(n - m) \qquad (\text{if } V_b > V_s \geq P \text{ and } m > n_b). \tag{5.63}$$

Next, suppose $V_b \geq P > V_s$. If $m \leq n_b$, then Case I or Case II holds, with $V_m = V_b$, $|N_m^\circ| = n_b$ and $|N_m^-| = n_s$; formulas (5.61) and (5.62) also apply to this situation. If $m > n_b$, then $V_m = V_s$ and Case III holds, with $|T \cup E| = n_b$, $|L| = 0$ and $|U| = |N_m^\circ| = n_s$. Case IIIA is characterized by $P(m - n_b) \geq V_s n_s$, in which case (5.31) and (5.32) yield

$$F^\circ = (V_b - P)n_b, \qquad G^\circ = 0 \qquad (\text{Case IIIA}). \tag{5.64}$$

Case IIIB is characterized in the present context by $P(m - n_b) \leq V_s n_s$; it has $\alpha = \beta = 0$, $\Delta = \lambda = m - n_b$ and $|B_\alpha|V_\alpha^*/P = V_s n_s/P$, so that (5.51) yields

$$F^\circ = V_{\text{tot}} - Pm \qquad (\text{Case IIIB}) \tag{5.65}$$

and (5.52) yields

$$G^\circ = -V_s(n - m) \qquad \text{if } P(m - n_b) < V_s n_s \qquad (\text{Case IIIB}), \tag{5.66}$$

$$-V_s(n-m) \le G^\circ \le 0 \qquad \text{if } P(m-n_b) = V_s n_s \qquad \text{(Case IIIB).} \tag{5.67}$$

Finally, suppose $P > V_b > V_s$, so that Case III applies with $|T \cup E| = 0$. If $m \le n_b$, then $|L| = 0$, $|N_m^\circ| = n_b$, $\beta = 1$ and $|B_1| = n_s$. Case IIIA is characterized by $Pm \ge V_{tot}$, and by (5.31)-(5.32) it yields $F^\circ = G^\circ = 0$. Case IIIB is characterized by $Pm \le V_{tot}$, and has $\Delta = m$. There are two subcases. First, if $Pm \le V_b n_b$ then $\alpha = 0$, $\lambda = m$, and (5.51)-(5.52) yield $F^\circ = V_{tot} - Pm$ together with

$$G^\circ = -(V_{tot} - V_b m) \qquad \text{(if } Pm < V_b n_b\text{)},$$

$$-(V_{tot} - V_b m) \le G^\circ \le -V_s(n-m) \qquad \text{(if } Pm = V_b n_b\text{)}.$$

Second, if $V_b n_v < Pm \le V_{tot}$, then $\alpha = 1$, $\lambda = m - V_b n_b/P$, and (5.51)-(5.52) yield $F^\circ = V_{tot} - Pm$ together with

$$G^\circ = -V_s(n-m) \qquad \text{(if } Pm < V_{tot}\text{)},$$

$$-V_s(n-m) \le G^\circ \le 0 \qquad \text{(if } Pm = V_{tot}\text{)}.$$

Still supposing $P > V_b > V_s$, now assume $m > n_b$. Then $|L| = n_b$, $|N_m^\circ| = n_s$ and $\alpha = \beta = 0$. Case IIIA is characterized by $Pm \ge V_{tot}$, and as above yields $F^\circ = G^\circ = 0$. Case IIIB is characterized by $Pm \le V_{tot}$, has $\Delta = \lambda = m - V_b n_b/P$; (5.51)-(5.52) yield $F^\circ = V_{tot} - Pm$ together with

$$G^\circ = -V_s(n-m) \qquad \text{(if } Pm < V_{tot}\text{)},$$

$$-V_s(n-m) \le G^\circ \le 0 \qquad \text{(if } Pm = V_{tot}\text{)}.$$

The dependence of F° upon P and m is the same as in Model 1, portrayed in the earlier figure 2. The dependence of G° is summarized in figure 4. G° is continuous on the horizontal boundary segment (of the line $m = n_b$), but is neither continuous across nor single-valued along the three hyperbolic boundary arcs. Note the presence of two distinct regions exhibiting penalty-saturation. (See table 2.)

TABLE 2. *Results for Model 2*

	Case Definition	Inspectee's Strategy	Inspector's Strategy	Payoffs
Case I	$V_m > P$	$c_i^\dagger = 1$ for all i	$p_i^\dagger = 1$ for $i \in N_m^+$ $p_i^\dagger = 0$ for $i \in N_m^-$ $0 < p_i^\dagger < 1$ for $i \in N_m^\circ$, $\sum\{p_i^\dagger: \ i \in N_m^\circ\} = m - \lvert N_m^+\rvert$	$F^\circ = V_{tot} - Pm$ $-G^\circ = [V_m\{n - \lvert N_m^-\rvert - M\} +$ $\sum\{V_i: \ i \in N_m^-\}]$
Case IIA	$V_m = P$ $\lvert N_m^+\rvert + \lvert N_m^\circ\rvert > m$	$c_i^\dagger = 1$ for all i	$p_i^\dagger = 1$ for $i \in N_m^+$ $p_i^\dagger = 0$ for $i \in N_m^-$ $0 < p_i^\dagger < 1$ for $i \in N_m^\circ$ $\sum\{p_i^\dagger: \ i \in N_m^\circ\} = m - \lvert N_m^+\rvert$	$F^\circ = V_{tot} - Pm$ $-G^\circ = [V_m\{n - \lvert N_m^-\rvert - m\} +$ $\sum\{V_i: \ i \in N_m^-\}]$
Case IIB	$V_m = P$ $\lvert N_m^+\rvert + \lvert N_m^\circ\rvert = m$	$c_i^\dagger = 1$ for $i \in N_m^+ = T$ $c_i^\dagger = 1$ for $i \in N_m^- = U$ $c_i^\dagger \geq V_{m+1}/P$ for $i \in N_m^\circ = E$	$p_i^\dagger = 1$ for $i \in N_m^+ \cup N_m^\circ = T \cup E$ $p_i^\dagger = 0$ for $i \in N_m^- = U$	$F^\circ = V_{tot} - Pm$ $-G^\circ = \sum\{V_i: \ i \in N_m^-\}]$
Case IIIA	$V_m < P$ $m \geq \lvert T \cup E\rvert + \sum\{V_i/P:$ $i \in U\}$	$c_i^\dagger = 1$ for $i \in T$ $0 \leq c_i^\dagger \leq, 1$ for $i \in E$ $c_i^\dagger = 0$ for $i \in U$	$p_i^\dagger = 1$ for $i \in T \cup E$ $p_i^\dagger \geq V_i/P$ for $i \in U$	$F^\circ = \sum\{V_i: \ i \in T\} - P\lvert T\rvert$ $G^\circ = 0$
Case IIIB	$V_m < P$ $m \leq \lvert T \cup E\rvert + \sum\{V_i/P:$ $i \in U\}$	$c_i^\dagger = 1$ for $i \in T$ $c_i^\dagger \geq c_m^\circ V_m/P$ for $i \in E$ $c_i^\dagger = c_m^\circ V_m/P$ for $i \in L$ $c_i^\dagger = c_m^\circ V_m/V_i$ for $\begin{cases} i \in B_b \text{ with } 0 \leq b < \alpha \\ i \in B_a \text{ if } \lambda = \lvert B_a\rvert V_a^*/P \end{cases}$ $c_i^\dagger = 1$ for $\begin{cases} i \in B_a \text{ if } \lambda < \lvert B_a\rvert V_a^*/P \\ i \in B \text{ with } \alpha < b \leq \beta \end{cases}$	$p_i^\dagger = 1$ for $i \in T \cup E$ $p_i^\dagger = V_i/P$ for $i \in L$ $p_i^\dagger = V_i/P$ for $\begin{cases} i \in B_b \text{ with } 0 \leq b < \alpha \\ i \in B_a \text{ if } \lambda = \lvert B_a\rvert V_a^*/P \end{cases}$ $p_i^\dagger \leq V_i/P$ for $i \in B_a$ if $\lambda < \lvert B_a\rvert V_a^*/P$ $p_i^\dagger = 0$ for $i \in B_a$ with $\alpha < b \leq \beta$	$F^\circ = V_{tot} - Pm$ $-G^\circ = \sum_{b>a}\lvert B_a\rvert V_a^* +$ $\{n - m - \sum_{b>a}\lvert B_a\rvert\}c_m^\circ V_m$

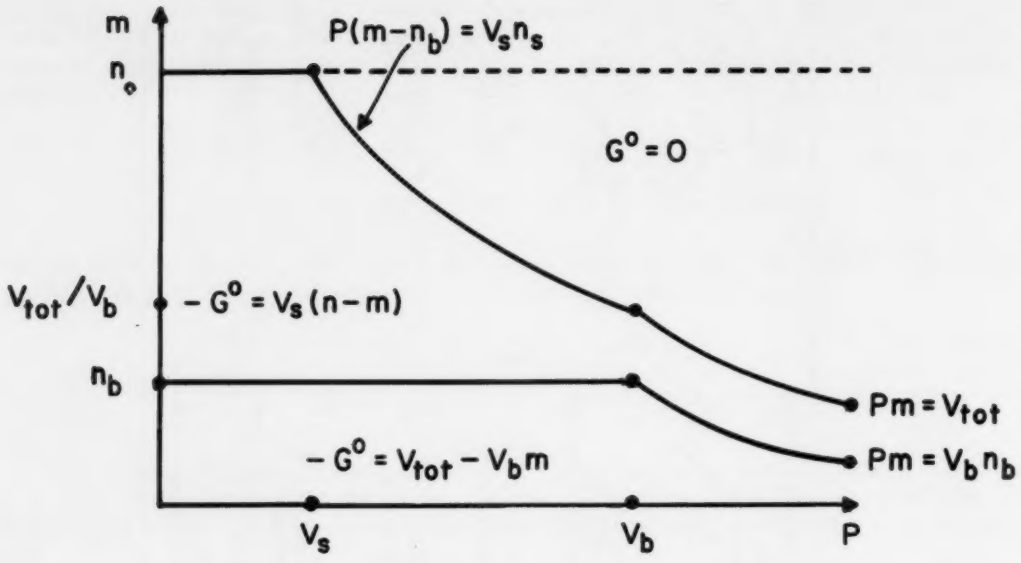


FIGURE 4. *Undetected illicit activity (big firms, small firms).*

6. Analysis for Model 2

The aim in this section is to prove the results cited in section 5, i.e., to determine the equilibrium points $(\mathbf{c}^o,\mathbf{p}^o)$ of the game with payoff functions

$$F(\mathbf{c},\mathbf{p}) = \sum_1^n [V_i - Pp_i]c_i, \qquad G(\mathbf{c},\mathbf{p}) = -\sum_1^n V_i c_i(1-p_i). \tag{6.1}$$

Mixed strategies for the inspectee (i.e., $0 \le c_i \le 1$) are permitted. The convention

$$V_1 \ge V_2 \ge \ldots \ge V_n \tag{6.2}$$

will remain in force, as will the definitions given in section 5 of the sets T, $\bar{T}$, E, U, N_m, N_m^+, N_m^o, and N_m^- and L. Of the two conditions defining an equilibrium point, (5.3) is equivalent to the pair of requirements

$$c_i^o = 1 \qquad \text{if } V_i > Pp_i^o, \tag{6.3}$$

$$c_i^o = 0 \qquad \text{if } V_i < Pp_i^o, \tag{6.4}$$

while satisfying (5.4) is equivalent to choosing $\mathbf{p}^o$ so as to achieve

$$\max_{\mathbf{p}} \sum_1^n (V_i c_i^0) p_i \qquad \text{subject to (2.3).} \tag{6.5}$$

An immediate consequence of (6.3) is that

$$c_i^0 = 1 \qquad \text{for } i \in T. \tag{6.6}$$

Another preliminary observation is that

$$|N_m^+| + |N_m^0| \geq m, \tag{6.7}$$

since $N_m^+ \cup N_m^0$ contains N_m.

We begin with the situation

$$V_m > P \qquad \text{(Case I).} \tag{6.8}$$

Here $N_m^+ \cup N_m^0$ is a subset of T, so that (6.6) yields

$$c_i^0 = 1 \qquad \text{for } i \in N_m^+ \cup N_m^0, \tag{6.9}$$

and since N_m^+ contains the $|N_m^+|$ largest V_i's and $|N_m^+| < m$, it follows from (6.5) that

$$p_i^0 = 1 \qquad \text{for } i \in N_m^+. \tag{6.10}$$

Combining (6.7), (6.9) and (6.5), it follows that

$$\sum \{p_i^0: \ i \in N_m^0\} = m - |N_m^+| \tag{6.11}$$

$$p_i^0 = 0 \qquad \text{for } i \in N_m^-. \tag{6.12}$$

By (6.3) and (6.12),

$$c_i^0 = 1 \qquad \text{for } i \in N_m^-. \tag{6.13}$$

Conversely, any $(\mathbf{c}^0, \mathbf{p}^0)$ which satisfies (6.9–6.13) also satisfies (6.3) through (6.5), and so is an equilibrium point. This completes the analysis of Case I.

Next assume

$$V_m = P \qquad \text{(Case II),} \tag{6.14}$$

so that N_m^+, N_m^0, N_m^- coincide with T, E, U respectively. By (6.6),

$$c_i^0 = 1 \qquad \text{for } i \in N_m^+, \tag{6.15}$$

so that (6.5) implies

$$p_i^0 = 1 \qquad \text{for } i \in N_m^+. \tag{6.16}$$

Define the set

$$N^{\circ}_{m1} = \{i: \quad i \in N^{\circ}_{m},\ c^{\circ}_{i} = 1\}; \tag{6.17}$$

we note that for $i \in N^{\circ}_{m} - N^{\circ}_{m1}$, (6.3) implies $0 \geq P - Pp^{\circ}_{i}$, *i.e.*,

$$p^{\circ}_{i} = 1 \qquad \text{for } i \in N^{\circ}_{m} - N^{\circ}_{m1}. \tag{6.18}$$

First assume that [1]

$$|N^{+}_{m}| + |N^{\circ}_{m1}| > m \qquad \text{(Case IIA)}. \tag{6.19}$$

Then (6.5) implies

$$p^{\circ}_{i} = 0 \qquad \text{for } i \in N^{\circ}_{m} - N^{\circ}_{m1}, \tag{6.20}$$

$$p^{\circ}_{i} = 0 \qquad \text{for } i \in N^{-}_{m}. \tag{6.21}$$

Comparison of (6.18) and (6.20) shows that $N^{\circ}_{m} - N^{\circ}_{m1}$ is empty, i.e.,

$$c^{\circ}_{i} = 1 \qquad \text{for } i \in N^{\circ}_{m}, \tag{6.22}$$

while (6.21) and (6.3) imply

$$c^{\circ}_{i} = 1 \qquad \text{for } i \in N^{-}_{m}, \tag{6.23}$$

and (6.16) and (6.20) imply

$$\sum \{p^{\circ}_{i}: \quad i \in N^{\circ}_{m}\} = m - |N^{+}_{m}|. \tag{6.24}$$

Conversely, such a $(\mathbf{c}^{\circ},\mathbf{p}^{\circ})$, which has the same structure as in Case I, satisfies (6.3–6.5) and so is an equilibrium point. The defining condition (6.19) takes the form

$$|N^{+}_{m}| + |N^{\circ}_{m}| > m, \tag{6.25}$$

i.e., m is not the last member of N°_{m}.

Next assume that

$$|N^{+}_{m}| + |N^{\circ}_{m1}| \leq m \qquad \text{(Case IIB)}. \tag{6.26}$$

Then (6.5) together with (6.18) imply

$$p^{\circ}_{i} = 1 \qquad \text{for } i \in N^{\circ}_{m}. \tag{6.27}$$

From this and (6.16) it follows that

$$m = \sum_{1}^{n} p^{\circ}_{i} \geq |N^{+}_{m}| + |N^{\circ}_{m}|,$$

and comparison with (6.7) shows that

$$|N^{+}_{m}| + |N^{\circ}_{m}| = m \tag{6.28}$$

[1] The analysis which follows will show that this kind of correspondence coincides with Case IIA as defined in section 5 (5.15). The same situation will recur in later subcases.

(hence there is no overlap with Case IIA's range) and that

$$p_i^\circ = 0 \qquad \text{for } i \in N_{\bar{m}}. \tag{6.29}$$

By (6.29) and (6.3)

$$c_i^\circ = 1 \qquad \text{for } i \in N_{\bar{m}}. \tag{6.30}$$

From (6.5) it follows, in view of (6.27) and (6.29), that

$$V_i c_i \geq V_j c_j \qquad \text{for } i \in N_m^\circ \text{ and } j \in N_{\bar{m}},$$

which by (6.28) and (6.30) is equivalent to

$$c_i \geq V_{m+1}/P \qquad \text{for } i \in N_m^\circ. \tag{6.31}$$

Conversely, if (6.28) holds, then any $(\mathbf{c}^\circ, \mathbf{p}^\circ)$ satisfying (6.15–6.16), (6.27) and (6.29–6.31) obeys (6.3–6.5) and so is an equilibrium point. This completes the analysis for Case II.

We turn now to the situation

$$V_m < P \qquad \text{(Case III)}. \tag{6.32}$$

By (6.5) and (6.6),

$$p_i^\circ = 1 \qquad \text{for } i \in T. \tag{6.33}$$

For $i \in E$, if $p_i^\circ < 1$ then (6.3) would imply $c_i^\circ = 1$, which by (6.5) would imply $p_i^\circ = 1$; hence

$$p_i^\circ = 1 \qquad \text{for } i \in E. \tag{6.34}$$

According to (5.34), U has been partitioned as $L \cup N_m^\circ \cup N_{\bar{m}}$, where

$$L = \{i: \ V_m < V_i < P\}; \tag{6.35}$$

Also, N_m^+ is partitioned as $T \cup E \cup L$. For $i \in L$, if $c_i^\circ = 1$ then (6.5) implies $p_i^\circ = 1$, which by (6.4) implies $c_i^\circ = 0$. Hence

$$c_i^\circ < 1 \qquad \text{for } i \in L, \tag{6.36}$$

which by (6.3) implies

$$p_i^\circ \geq V_i/P \qquad \text{for } i \in L. \tag{6.37}$$

For each of $\mathbf{c}^\circ$, let

$$M = \max \{V_i c_i^\circ: \ i \in U\}, \tag{6.38}$$

$$S = \{i: \ i \in U, V_i c_i^\circ = M\}. \tag{6.39}$$

Then (6.34) and (6.5) imply

$$c_i^\circ \geq M/P \qquad \text{for } i \in E. \tag{6.40}$$

If $|S| \leq m - |T \cup E|$, then (6.5) would imply $p_i^\circ = 1$ for $i \in S$, in which case (6.4) implies $c_i^\circ = 0$ for $i \in S$, so that $M = 0$ and $S = U$, which (since $m < n$) would contradict the supposition $|S| \leq m - |T \cup E|$. Hence

$$|T \cup E| + |S| > m. \tag{6.41}$$

By (6.5) followed by (6.3),

$$p_i^\circ = 0 \qquad \text{for } i \in U - S, \tag{6.42}$$

$$c_i^\circ = 1 \qquad \text{for } i \in U - S. \tag{6.43}$$

It follows from (6.43) and the definition of M that

$$M > \max \{V_i: \ i \in U - S\}; \tag{6.44}$$

in particular, for each $s \in S$, V_s exceeds that maximum. An immediate consequence is that

$$L \text{ is contained in } S. \tag{6.45}$$

By (6.41), S cannot consist just of L, and then (6.44) implies

$$N_m^\circ \text{ is contained in } S. \tag{6.46}$$

In particular $m \in S$, so that

$$M = V_m c_m^\circ. \tag{6.47}$$

At this point it is convenient to consider separately the situation

$$M = 0, \qquad \text{(Case IIIA)} \tag{6.48}$$

or equivalently

$$c_i^\circ = 0 \qquad \text{for } i \in U, \tag{6.49}$$

which by (6.3) requires

$$p_i^\circ \geq V_i/P \qquad \text{for } i \in U, \tag{6.50}$$

implying

$$m \geq |T \cup E| + \sum \{V_i/P: \ i \in U\}. \tag{6.51}$$

Conversely, under Case III, if (6.51) holds, then there exist pairs $(\mathbf{c}^\circ, \mathbf{p}^\circ)$ satisfying (6.6), (6.49), (6.33–6.34) and (6.50); since any such pair obeys (6.3–6.5), it is an equilibrium point.

From now on, assume

$$M > 0 \qquad \text{(Case IIIB)}. \tag{6.52}$$

Then it follows from (6.45) that $c_i^\circ > 0$ for $i \in L$, so that (6.4) implies $p_i^\circ \leq V_i/P$, which together with (6.37) yields

$$p_i^\circ = V_i/P \qquad \text{for } i \in L. \tag{6.53}$$

A similar argument using (6.46) yields

$$p_i^\circ \leq V_m/P \qquad \text{for } i \in N_m^\circ. \tag{6.54}$$

Also, (6.45) implies

$$c_i^\circ = V_m c_m^\circ / V_i \qquad \text{for } i \in L, \tag{6.55}$$

while (6.46) gives

$$c_i^\circ = c_m^\circ \qquad \text{for } i \in N_m^\circ. \tag{6.56}$$

Combining (6.3) with (6.54) yields

$$p_i^\circ = V_m/P \qquad \text{for } i \in N_m^\circ, \text{ if } c_m^\circ < 1. \tag{6.57}$$

Now partition N_m^- into β blocks B_b, as described following (5.34). Then by (6.44), $S \cap N_m^-$ consists of the first α of these blocks, for some integer α with $0 \geq \alpha \geq \beta$. For each block B_b in S, $\{c_i^\circ: \ i \in B_b\}$ consists of a single number c_b^*, and

$$c_b^* = c_m^\circ V_m / V_b^* \qquad \text{for } 1 \leq b \leq \alpha. \tag{6.58}$$

Since $c_b^* \leq 1$, (6.58) yields

$$V_\alpha^* \geq c_m^\circ V_m. \tag{6.59}$$

On the other hand, if $\alpha < \beta$, then (6.44) and (6.47) imply

$$c_m^\circ > V_{\alpha+1}^*/P \qquad \text{(if } \alpha < \beta\text{)}. \tag{6.60}$$

If $c_m^\circ = 1$, this implies $\alpha = 0$, and requires

$$|T \cup E| + \sum\{V_i/P: \ i \in L\} + |N_m^\circ| V_m/P \geq m. \tag{6.61}$$

If $c_m^\circ < 1$, then (6.58) implies $0 < c_b^* < 1$ for $1 \leq b < \alpha$, and thus by (6.3) and (6.4)

$$p_i^\circ = V_b^*/P \qquad \text{for } i \in B_b \text{ and } 1 \leq b < \alpha. \tag{6.62}$$

Since $c^* \leq 1$, (6.4) yields

$$p_i^\circ \geq V_\alpha^*/P \qquad \text{for } i \in B_\alpha. \tag{6.63}$$

It follows from these results that

$$\sum_{b=0}^{\alpha} |B_b|V_b/P \geq m - |T \cup E| - \sum\{V_i/P: \quad i \in L\} \geq \sum_{b=0}^{\alpha-1} |B_b|V_b^*/P. \tag{6.64}$$

Equality in the second relation is equivalent to equality in (6.63); this situation corresponds in the notation of (5.37) to $\lambda = |B_\alpha|V_\alpha^*/P$. If on the other hand $\lambda < |B_\alpha|V_\alpha^*/P$, then since

$$\sum\{p_i^0: \quad i \in B_\alpha\} = \lambda \tag{6.65}$$

by virtue of (6.33–6.34), (6.42), (6.52), (6.57) and (6.60), it follows that $p_i^0 < V_\alpha^*/P$ must hold for some $i \in B_\alpha$, and so by (6.3) we have $c_\alpha^* = 1$, which by (6.48) fixes the value of c_m^0 at V_α^*/V_m^*.

Finally, observe that (6.51) and (6.64) will hold simultaneously if and only if the former holds as an equality while the latter's second part holds as an equality with $\alpha = \beta$.

7. Formulation and Results for Model 3

Recall that section 5 originally set out to analyze the consequences of the payoff functions

$$F(\mathbf{c},\mathbf{p}) = \sum_1^n [V_i - Pp_i]c_i, \tag{7.1}$$

$$G(\mathbf{c},\mathbf{p}) = -\sum_1^n V_i c_i \tag{7.2}$$

for inspectee and inspector respectively, but that this was found unfruitful within the framework of "equilibrium point" solutions to nonzero-sum games. The response in section 5 was to replace (7.2) with an alternative function (5.5) representing the loss (e.g., to consumers) due to undetected cheating (rather than all cheating). In this section we explore a different approach, in which (7.2) is retained but the customary game-theoretic framework is altered.

Specifically, the critical assumption here is that no matter what strategy $\mathbf{p}$ is selected by the inspector, the inspectee learns of it in advance (or can estimate it through experience), and so is able to select a strategy $\mathbf{c}(\mathbf{p})$ which maximizes his payoff, i.e.,

$$\mathrm{F}[\mathbf{c}(\mathbf{p}),\mathbf{p}] = \max_{\mathbf{c}} F(\mathbf{c},\mathbf{p}). \tag{7.3}$$

(In general $\mathbf{p}$ is a mixed strategy, so that the inspectee need not know exactly which devices will be inspected but only the associated probabilities.) The inspector's problem, therefore, is to choose $\mathbf{p}$ so as to maximize the function

$$H(\mathbf{p}) = G[\mathbf{c}(\mathbf{p}),\mathbf{p}]. \tag{7.4}$$

To discuss this approach further, it is convenient to make the further hypothesis that each component of $\mathbf{c}(\mathbf{p})$ depends only on the corresponding component of $\mathbf{p}$, so that we can write $c_i(p_i)$ rather than $c_i(\mathbf{p})$. This hypothesis is reasonable *per se* (since the inspectee's c_i's are not linked by any constraints), and also is consistent with the consequences

$$c_i(p_i) = 0 \qquad \text{if } p_i > V_i/P, \tag{7.5}$$

$$c_i(p_i) = 1 \qquad \text{if } p_i < V_i/P \tag{7.6}$$

of (7.1) and (7.3).

In terms of the set[2]

$$E(\mathbf{p}) = \{i: \quad p_i = V_i/P\}, \tag{7.7}$$

we can identify a set of components of $\mathbf{c}(\mathbf{p})$, namely

$$\{c_i(p_i): \quad i \in E(\mathbf{p})\},$$

which are not determined by the criterion (7.3). That is, $\mathbf{c}(\mathbf{p})$ is not single-valued, and so the same is true of the inspector's objective function

$$H(\mathbf{p}) = -\sum_1^n V_i c_i(p_i). \tag{7.8}$$

Some further assumption is needed to resolve this ambiguity. For example, since our critical assumption (that the inspectee knows $\mathbf{p}$) is "pessimistic" from the inspector's viewpoint, one might wish to be consistently pessimistic and thus to replace the right-hand side of (7.4) by its minimum over all c(p) consistent with (7.3). This is equivalent to setting

$$c_i(p_i) = 1 \qquad \text{all } i \in E(\mathbf{p}).$$

At the opposite extreme, one might want to introduce a countervailing bias in the optimistic direction, and so would set

$$c_i(p_i) = 0 \qquad \text{all } i \in E(\mathbf{p}). \tag{7.9}$$

This is the choice that will for the present be made. It involves a notion of "no cheating without a positive expectation of gain," which is not implausible but certainly involves appeal to considerations that are external to the model (and that, in particular, go beyond the penalty P and its deterrent effect). After the consequences of hypothesis (7.9) have been analyzed, we will consider the more general situation in which this assumption is not imposed.

Retaining the notation

$$V_{\text{tot}} = \sum_1^n V_i,$$

define

$$V(I) = \sum\{V_i: \quad i \in I\}$$

for subsets I of N, and set

$$I(\mathbf{p}) = \{i: \quad p_i \geq V_i/P\}. \tag{7.10}$$

Then $H(\mathbf{p})$ depends on $\mathbf{p}$ only via $I(\mathbf{p})$, since by (7.5), (7.6), (7.8) and (7.9),

$$\begin{aligned} H(\mathbf{p}) &= -\sum\{V_i: \quad i \in N - I(\mathbf{p})\} \\ &= -V_{\text{tot}} + V\,[I(p)]. \end{aligned} \tag{7.11}$$

[2] The following notation should not be confused with the usage of symbol "E" in (3.10) or with the expected value function.

Note that $I(\mathbf{p})$ corresponds to the set of devices on which cheating will not occur, according to (7.5) and (7.9). From (7.11) it follows that the inspector's problem can be written

$$\text{choose } \mathbf{p} \text{ to maximize } V[I(\mathbf{p})]. \tag{7.12}$$

In this context, it is natural to ask which subsets I of N can arise in the form $I(\mathbf{p})$ for some $\mathbf{p}$, i.e., as the "non-cheating" set for some strategy of the inspector. As will be shown in section 8, a set of conditions on a subset I which are both necessary and sufficient that $I=I(\mathbf{p})$ for some $\mathbf{p}$, are

$$V(I) \leq Pm, \tag{7.13}$$

$$I \subseteq \bar{T} = N - T, \tag{7.14}$$

$$P(m - |I| - |T|) < V(N - I - T). \tag{7.15}$$

For a subset I of N which obeys these three conditions, there may be many choices of $\mathbf{p}$ for which $I(\mathbf{p}) = I$. If $|I| + |T| < m$, one such $\mathbf{p}$ is given by

$$\begin{aligned} p_i &= 1 && \text{for } i \in I \cup T \\ p_i &= [V_i(m - |I| - |T|)]/V(N - I - T) && \text{for } i \in N - I - T. \end{aligned} \tag{7.16}$$

If $|I| \leq m \leq |I| + |T|$, then such a $\mathbf{p}$ is given by

$$\begin{aligned} p_i &= 1 && \text{for } i \in I, \\ p_i &= [m - |I|]/|T| && \text{for } i \in T, \\ p_i &= 0 && \text{for } i \in N - I - T. \end{aligned} \tag{7.17}$$

If $|I| > m$ then such a $\mathbf{p}$ is given by

$$p_i = V_i/P + (1 - V_i/P)\,[Pm - V(I)]/[P|I| - V(I)] \qquad \text{for } i \in I, \tag{7.18}$$

$$p_i = 0 \qquad \text{for } i \in N - I. \tag{7.19}$$

(When $P|I| - V(I) = 0$ then $V_i = P$ for all $i \in I$ and we interpret (7.18) as $p_i = 1$.) This will be proved in section 8.

The inspector's problem can now be rephrased as that of choosing a subset I of N, subject to (7.13) through (7.15), to

$$\text{maximize } V(I). \tag{7.20}$$

When such an I has been found, an optimal $\mathbf{p}$ can be calculated through (7.16), (7.17) or (7.18–7.19). For the maximization problem, (7.15) and (7.16) are, in fact, unnecessary, that is, if I maximizes $V(I)$ subject to (7.13) and (7.14) then $|I| + |T| \geq m$, so that (7.16) does not arise, and I satisfies (7.15). For a proof, suppose that $|I| + |T| < m$. Since $|\bar{T}| + |T| = n > m$, it follows that I is a proper subset of $\bar{T}$. Let $j \in \bar{T}$, $j \notin I$, and set

$$J = I \cup \{j\}.$$

Since $|I| \leq |I| + |T| < m$, we have $|I| \leq m - 1$ and so

$$|J| \leq |J| + |T| = |I| + |T| + 1 \leq m.$$

Clearly J satisfies (7.13) and (7.14) and the left-hand side of (7.15) is nonpositive. The left-hand side of (7.15) is equal to zero only when $|J| + |T| = m$ and, in this case, J is a proper subset of $\bar{T}$, so that the right-hand side of (7.15) is positive. Hence J satisfies (7.15) as well. This contradicts the assumption that I maximizes $V(I)$ subject to (7.13—7.15).

For yet another formulation of the inspector's problem, introduce a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of discrete variables

$$x_i = \begin{cases} 1 \text{ for } i \in I \\ 0 \text{ for } i \in N - I. \end{cases}$$

Then the problem is to choose $\mathbf{x}$ to

$$\text{maximize } \sum_1^n V_i x_i \tag{7.21}$$

subject to the analogs of (7.13) and (7.14):

$$\textstyle\sum_1^n V_i x_i \leq Pm, \tag{7.22}$$

$$x_i = 0 \qquad \text{for } i \in T. \tag{7.23}$$

Problem (7.21)–(7.23) cannot in general be solved in closed form. As a (binary) integer program, it can (if n is not too large) be solved by any of the methods developed for such problems. More specifically, it is a "knapsack problem," for which special algorithms are available.[3] Its peculiarity as a knapsack problem, namely the presence of the same coefficients (V_i) in both (7.21) and (7.22), suggest the possibility of more efficient "tailored" methods, but this will not be pursued here.

EXAMPLE 1: *Equal-Sized Firms*. As before, we assume in this scenario that all $V_i = V$. If $P < V$, then T is all of N, so that (7.14) implies I is empty. Thus (7.17) applies, giving

$$p_i = m/n \qquad \text{for all } i,$$

and the associated payoffs are

$$F^0 = V_{\text{tot}} - Pm, \tag{7.24}$$

$$G^0 = -V_{\text{tot}}. \tag{7.25}$$

This case corresponds to weak social sanctions, resulting in cheating on every device.

If $P \geq V$, then T is empty and the problem is that of choosing I to maximize $|I|$ subject to

$$|I| \leq Pm/V. \tag{7.26}$$

To describe the solution in this case, let k be the largest integer not exceeding either Pm/V or n, and let I be *any* subset of N with $|I| = k$. Since $n > m$ and $P \geq V$, it follows that $k \geq m$, so (7.18–7.19)

[3] See, e.g., A.V. Cabot, "An Enumeration Algorithm for Knapsack Problems," Operations Research **18** (1970), 306–311.

applies to yield

$$p_i = m/k \qquad \text{for } i \in I,$$

$$p_i = 0 \qquad \text{for } i \in N - I.$$

The associated inspectee strategy is

$$c_i = 0 \qquad \text{for } i \in I,$$

$$c_i = 1 \qquad \text{for } i \in N - I,$$

and the resultant payoffs are

$$F^\circ = -G^\circ = V(n - k) = V_{\text{tot}} - Vk. \tag{7.27}$$

See figure 5 for a depiction of the payoffs a functions of P and m. F° is continuous along the line $P = V$, whereas G° has a discontinuity across that line. To the right of this line, F° and G° are both discontinuous across each hyperbola $Pm = tV$ for integers $t \leq n$, since the integer parameter k changes value there.

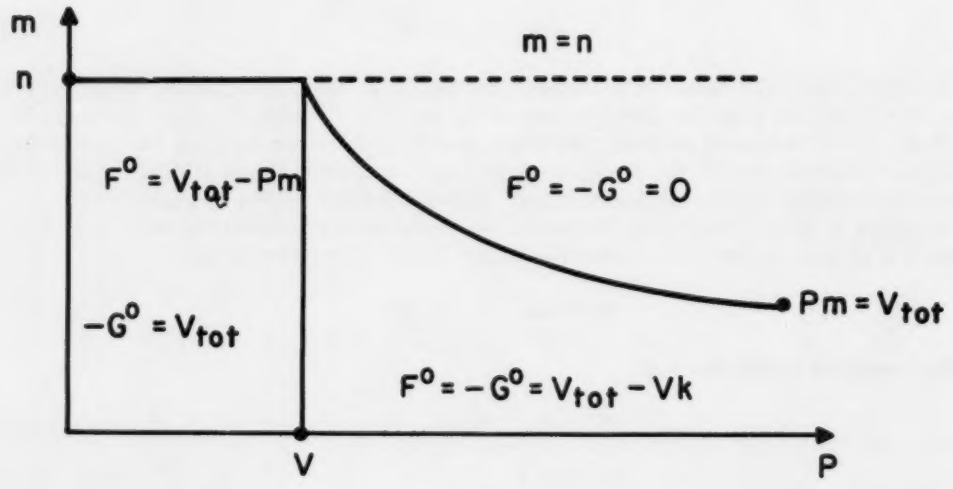


FIGURE 5. *Net illicit gain* (F°) *and total illicit-activity level* ($-G^\circ$). (Equal-sized firms) (k is greatest integer $\leq Pm/V$).

EXAMPLE 2: *Big Firms, Little Firms*. As before, N is partitioned into nonempty subsets, $N = B \cup S$, with

$$V_i = V_b \qquad \text{for } i \in B$$

$$V_i = V_s \qquad \text{for } i \in S,$$

and $V_b > V_s$. The notations $n_b = |B|$ and $n_s = |S|$ will again be used.

If $V_b > V_s > P$ then T is all of N, so that (7.14) implies I is empty (hence all $c_i = 1$). Thus (7.17) applies, giving

$$p_i = m/n \qquad \text{for } i \in N.$$

The corresponding payoffs are

$$F^\circ = V_{\text{tot}} - Pm, \qquad G^\circ = -V_{\text{tot}}. \tag{7.28}$$

If $V_b > P \geq V_s$, then $T = B$ and, by (7.6), $c_i = 1$ for all $i \in B$. Thus by (7.14), I must be a subset of S, and the inspector's problem is to choose such a subset to maximize $|I|$, subject to

$$|I| \leq Pm/V_s. \tag{7.29}$$

Two cases arise:

In case $n_s \leq m$, we take $I = S$. This is clearly the largest possible subset of S, and it satisfies (7.29) since $n_s \leq$ m and $V_s \leq P$. From (7.17)

$$p_i = 1 \qquad \text{for all } i \in S,$$

$$p_i = (m - n_s)/n_b \qquad \text{for all } i \in B,$$

and by (7.5) and (7.9), $c_i = 0$ for all $i \in S$. The resultant payoffs are

$$F^\circ = -G^\circ = V_b n_b. \tag{7.30}$$

In the remaining case $m < n_s$, let k be the largest integer not exceeding either Pm/V_s or n_s, and let I be any subset of S with $|I| = k$. (Thus $k = n_s$ and $I = S$ when $Pm \geq V_s n_s$.) Since $P \geq V_s$ (and hence $Pm/V_s \geq m$), and $n_s > m$, it follows that $k \geq m$, and so (7.18–7.19) applies to yield

$$p_i = m/k \qquad \text{for } i \in I,$$

$$p_i = 0 \qquad \text{for } i \in N - I.$$

Since $m/k \geq V/P$, it then follows from (7.5) and (7.9) that $c_i = 0$ for $i \in I$, while by (7.6), $c_i = 1$ for $i \in S - I$. The corresponding payoffs are

$$F^\circ = -G^\circ = V_{\text{tot}} - V_s k. \tag{7.31}$$

Finally, if $P \geq V_b > V_s$, then T is empty. Let $x_b = |I \cap B|$ and $x_s = |I \cap S|$. The inspector's problem is to choose integers x_b and x_s, which

$$\text{maximize } z = V_b x_b + V_s x_s \tag{7.32}$$

subject to

$$0 \leq x_b \leq n_b, \qquad 0 \leq x_s \leq n_s, \tag{7.33}$$

$$V_b x_b + V_s x_s \leq Pm. \tag{7.34}$$

If $Pm \geq V_{tot}$ then the optimal I is all of N (i.e., $x_b = n_b$ and $x_s = n_s$), so that all $c_i = 0$, implying $F^o = G^o = 0$, while (7.18) gives, for all i,

$$p_i = V_i/P + (1 - V_i/P)[(Pm - V_{tot})/(Pn - V_{tot})].$$

Now assume $Pm < V_{tot}$. No closed-form answer to (7.32–7.34) seems possible, but trial-and-error solution methods should suffice. The maximum value z_{max} in (7.32) can be estimated by

$$Pm - V_b < z_{max} \leq Pm; \tag{7.35}$$

here the right-hand inequality follows from (7.34), while the left-hand one follows by observing that if $z \leq Pm - V_b$ then either x_b can be increased by 1 (thus increasing z by V_b) without violating (7.33) or (7.34), or else $x_b = n_b$ so that (since $z \leq Pm < V_{tot}$) $x_s < n_s$, implying x_s can be increased by 1 without violating the constraints. (If $Pm \leq V_s n_s$, similar logic shows the left-hand side of (7.35) can be sharpened—i.e., increased—to $Pm - V_s$.) Furthermore, any optimal solution must satisfy $x_b + x_s \geq m$, since otherwise

$$z_{max} = V_b x_b + V_s x_s \leq V_b(x_b + x_s) \leq V_b(m - 1) \leq P(m - 1) = Pm - P < Pm - V_b,$$

contradicting (7.35). Thus $|I| \geq m$, so that (7.18–7.19) applies to yield

$$p_i = V_i/P + (1 - V_i/P)\,[Pm - z_{max}]/[P(x_b + x_s) - z_{max}] \tag{7.36}$$

for $I \in I$, and $p_i = 0$ otherwise. Thus $c_i = 0$ for $i \in I$ and $c_i = 1$ otherwise, and the associated payoffs are

$$F^o = -G^o = V_b(n_b - x_b) + V_s(n_s - x_s) = V_{tot} - z_{max}, \tag{7.37}$$

so that (7.35) yields the estimates

$$V_{tot} - Pm \leq F^o = -G^o < V_{tot} - Pm + V_b. \tag{7.38}$$

(In accordance with a parenthetical remark above, the term V_b in (7.38) can be reduced to V_s if $Pm \leq V_s n_s$.)

Figure 6 shows the payoffs F^o and G^o as functions of P and m. This concludes the discussion of Example 2 in the context of Model 3.

It remains to discuss the consequences of dropping assumption (7.9). For this purpose, we introduce the numbers

$$c_i^* = c_i(V_i/P), \tag{7.39}$$

which necessarily satisfy $0 \leq c_i^* \leq 1$, and also set

$$N_+ = \{i \in N: \ \ c_i^* > 0\}. \tag{7.40}$$

In practice, it is unlikely that the c_i^*'s (i.e., the probabilities of cheating at the various devices in a context of zero expected gain) would be known accurately. They might be assigned nominal high, medium or low values (such as 0.90, 0.50, and 0.10) in accordance with the insight or degree of optimism applicable to a specific application. They might well be assumed equal over all devices, or over each group of devices in a classification into a very few groups; these common values could then be varied in a parametric sensitivity analysis. At any rate, these numbers are treated as "given" in what follows.

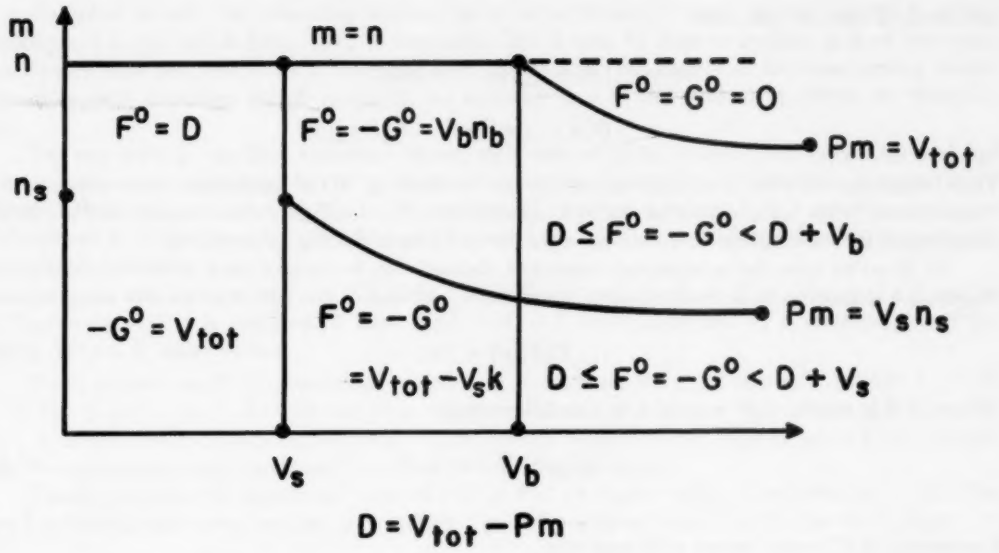


FIGURE 6. *Net illicit gain* (F^0) *and total illicit-activity level* ($-G^0$).
(Big firms, small firms) (k is greatest integer $\le Pm/V_s$) ($D = V_{tot} - Pm$).

In addition to the set $E(\mathbf{p})$ defined in (7.7), it is convenient to define further sets

$$M(\mathbf{p}) = \{i: \quad p_i > V_i/P\}, \tag{7.41}$$

$$L(\mathbf{p}) = \{i: \quad P_i < V_i/P\}, \tag{7.42}$$

so that $M(\mathbf{p})$, $E(\mathbf{p})$ and $L(\mathbf{p})$ form a partition of N. By (7.5) and (7.6),

$$\begin{aligned} c_i &= 0 \qquad \text{for } i \in M(\mathbf{p}), \\ c_i &= 1 \qquad \text{for } i \in L(\mathbf{p}). \end{aligned} \tag{7.43}$$

while by (7.39)

$$c_i = c_i^* \qquad \text{for } i \in E(\mathbf{p}). \tag{7.44}$$

It follows that

$$-H(\mathbf{p}) = \sum \{V_i c_i^*: \quad i \in E(\mathbf{p}) \cap N_+\} + V[L(\mathbf{p})]. \tag{7.45}$$

The members of the set $E(\mathbf{p}) \cap N_+$ constitute the "violations" of the previous assumption (7.9). At this point, we temporarily restrict attention to the special case in which the set E of (3.10) is empty, i.e., no V_i is exactly equal to P. Then for $i \in E(\mathbf{p})$, we have $p_i < 1$. In this case, as will be shown in section 8, such violations cannot occur for an *optimal* $\mathbf{p}$ unless the two conditions

$$M(\mathbf{p}) \text{ is empty,} \tag{7.46}$$

$$p_i = 0 \qquad \text{for all } i \in L(\mathbf{p}) \tag{7.47}$$

both hold. If they do hold, then

$$m = \sum_1^n p_i = \sum \{p_i: \quad i \in E(\mathbf{p})\}$$

$$= \sum \{V_i/P: \quad i \in E(\mathbf{p})\} = V[E(\mathbf{p})]/P.$$

Thus violations of (7.9) can only occur for an optimal $\mathbf{p}$ if the problem's data present the "coincidence" that some subset of the V_i's, specifically $\{V_i: \quad i \in E(\mathbf{p})\}$, sums exactly to Pm. In all other cases, the assumption (7.9) does not alter the problem of finding an optimal $\mathbf{p}$.

To describe how the exceptional cases just defined can be treated (and identified), define an inspector's strategy $\mathbf{p}$ to be *exceptional* if it satisfies (7.46) and (7.47), which as already seen, implies

$$V[E(\mathbf{p})] = Pm. \tag{7.48}$$

Since $i \in E(\mathbf{p})$ implies $V_i/P = p_i \leq 1$, it also follows that

$$E(\mathbf{p}) \subseteq \bar{T} = N - T. \tag{7.49}$$

Conversely, if E^* is any subset of N such that

$$V(E^*) = Pm, \qquad E^* \subseteq \bar{T}, \tag{7.50}$$

then $E^* = E(\mathbf{p})$ for one (in fact, precisely one) exceptional $\mathbf{p}$, namely the one defined by $p_i = V_i/P$ for $i \in E^*$ and all other $p_i = 0$. For exceptional $\mathbf{p}$'s, (7.45) becomes

$$-H(\mathbf{p}) = \sum \{V_i c_i^*: \quad i \in E(\mathbf{p})\} + V\,[n - E(\mathbf{p})]$$

$$= \sum \{V_i c_i^*: \quad i \in E(\mathbf{p})\} + V(N) - V[E(\mathbf{p})]$$

$$= \sum \{V_i c_i^*: \quad i \in E(\mathbf{p})\} + V_{\text{tot}} - Pm,$$

so that the problem of selecting a "best" exceptional $\mathbf{p}$ is equivalent to that of choosing a subset E^* of N to

$$\text{minimize} \sum \{V_i c_i^*: \quad i \in E^*\} \tag{7.51}$$

subject to (7.50). This can be written as a binary integer program in terms of variables

$$y_i = \begin{cases} 1 \text{ if } i \in E^*, \\ 0 \text{ otherwise,} \end{cases}$$

namely

$$\text{minimize} \sum_1^n V_i c_i^* y_i \tag{7.52}$$

subject to

$$\sum_1^n V_i y_i = Pm, \tag{7.53}$$

$$y_i = 0 \qquad \text{if } i \in T. \tag{7.54}$$

As noted already, the preceding analysis leads (in particular) to those optimal $\mathbf{p}$'s which exhibit violations, i.e., for which $E(\mathbf{p}) \cap N_+$ is nonempty. But it may be that no optimal $\mathbf{p}$ is of this type, and in any case the restriction to this type may unduly limit the variety of alternate optima found. Therefore, still assuming that E is empty, we consider how to find optimal $\mathbf{p}$ which are violation-free.

For any such $\mathbf{p}$, the first summand on the right side of (7.45) is zero, and since $L(\mathbf{p})$ and $I(\mathbf{p})$ are complementary subsets of N, the problem becomes that of choosing violation-free $\mathbf{p}$ to maximize $V[I(\mathbf{p})]$. This leads, as below (7.12), to the problem of choosing a subset I of N so as to maximize $V(I)$ subject to (7.13) through (7.15). As noted below (7.20), (7.15) is superfluous for this purpose and (7.16) does not arise.

Suppose that $|I| \leq m \leq |I| + |T|$. Then, by (7.17), no $i \in I$ can lie in $E(\mathbf{p})$. Also, no $i \in T$ can lie in $E(\mathbf{p})$ since $p_i \leq 1$ is required. If some $i \in N - I - T$ lay in $E(\mathbf{p})$ then $V_i = 0$, since $p_i = 0$ by (7.17). Thus $\mathbf{p}$ is violation-free.

Next, suppose such an optimizing I satisfies $|I| > m$. Since the emptiness of E implies $1 - V_i/P \neq 0$, the $\mathbf{p}$ given by (7.18–7.19) has $E(\mathbf{p})$ empty and hence is violation-free, *except* if $V(I) = Pm$ (i.e., if equality holds in (7.13)). So if $V(I) < Pm$, then a comparison of $H(\mathbf{p})$ as given by (7.11) with the "exceptional-strategy maximum" of H will identify the optimum.

Finally, suppose the maximum value of $V(I)$ is Pm. (A higher value is forbidden by (7.13).) For any I achieving this value, and any $\mathbf{p}$ for which $I(\mathbf{p}) = I$, we have (since $V_i \leq Pp_i$ for all $i \in I(\mathbf{p})$)

$$\begin{aligned} V[I(\mathbf{p})] &= \sum\{V_i\colon \ i \in I(\mathbf{p})\} \\ &\leq P \sum\{p_i\colon \ i \in I(\mathbf{p})\} \leq P \sum_1^n p_i = Pm, \end{aligned} \tag{7.55}$$

and since the end terms in (7.55) are equal, equality must hold throughout, implying that $\mathbf{p}$ is given by

$$p_i = V_i/P \qquad \text{for } i \in I, \tag{7.56}$$

$$p_i = 0 \qquad \text{for } i \in N - I. \tag{7.57}$$

This $\mathbf{p}$ is exceptional, and will exhibit violations unless $I \cap N_+$ is empty (for example, violations will occur if all $c_i^* > 0$). If $I \cap N_+$ is nonempty for all I obeying (7.13–7.15) and yielding $V(I) = Pm$, then finding a best violation-free $\mathbf{p}$ is not constrained by sharpening (7.13) to

$$V(I) < Pm. \tag{7.13'}$$

In this case, for any I maximizing $V(I)$, it follows as before that the $\mathbf{p}$ given by (7.16), (7.17) or (7.18-7.19) is violation-free and hence is a best violation-free strategy for the inspector.

In the context of the last paragraph, it is necessary to check whether the constraint (7.15) can still be omitted in maximizing $V(I)$. If $Pm < V(\bar{T})$ then the justifying argument given below (7.20) remains valid. The same is true when $Pm \geq V(\bar{T})$, except if $V(\bar{T}) = V(U) = Pm$. In this case (7.13) will be replaced by (7.13') and the solutions will be all sets $I = \bar{T} - \{j\}$, where j is such that $V_j = \min\{V_i\colon \ i \in \bar{T}\}$. Since $Pm = V(\bar{T})$, (7.15) is equivalent to $V(I) < P|I| + P|T|$, which is true when $|I| \neq 0$ (since $I \subseteq \bar{T}$) and is also true when $|I| = 0$ (since $|T| > 0$).

EXAMPLE 1': *Equal-Sized Firms*. Here we illustrate the preceding material. The scenario (all $V_i = V$) treated in Example 1 of this section is retained. However, the assumption (7.9) is no longer retained. It is assumed that E is empty, i.e., $V \neq P$. "Violations" of (7.9) can occur only if $Pm = kV$ for some integer k with $k \leq n$. When no such k exists, the analysis of Example 1 remains valid.

For the present example, we therefore assume that $Pm = kV$ where k is an integer with $k \leq n$. It is also necessary to specify the numbers c_i^* of (7.39). Without loss of generality, the devices may

be supposed numbered so that

$$0 \leq c_1^* \leq c_2^* \leq \ldots \leq c_n^* \leq 1. \tag{7.58}$$

If $P < V$, then T is all of N, so (7.54) implies that E^* is empty. Thus no exceptions to (7.9) arise, so that the analysis in Example 1 is applicable.

Now suppose $P > V$. First a best exceptional strategy, $\mathbf{p}^E$, will be determined. Since T is empty, (7.54) is vacuous. Since $Pm/V = k$, problem (7.52–7.53) becomes that of choosing a subset E^* of N to minimize $\sum\{c_i^*: \ i \in E^*\}$ subject to $|E^*| = k$. Clearly a solution is given by $E^* = \{1,2, \ldots, k\}$, and by the remark following (7.50), $\mathbf{p}^E$ is given by

$$p_i^E = V/P \qquad \text{for } i \in E^*, \qquad p_i^E = 0 \text{ otherwise.}$$

Use of the formula preceding (7.51) gives

$$H(\mathbf{p}^E) = -V_{\text{tot}} + Pm - V \sum_1^k c_i^*. \tag{7.59}$$

Now a best violation-free strategy, $\mathbf{p}^F$, will be determined. The problem of maximizing $V(I)$, subject to (7.13–7.14), has as solutions all subsets I of N with $|I| = k$, implying $V(I) = kV = Pm$. There are two subcases. If I can be chosen disjoint from N_+, which is true if and only if $c_k^* = 0$ in (7.58), then in particular I can be taken as the set E^* given above; here $\mathbf{p}^E$ is violation-free, can be taken as $\mathbf{p}^F$, and so is optimal. But if $c_k^* > 0$, then (7.13) must be replaced by (7.13′) in the maximization of $V(I)$. This problem has as solution any subset I of N with $|I| = k - 1$ (note that $P > V$ and $Pm = kV$ imply $k > m \geq 1$), in particular $I = \{1,2, \ldots, k-1\}$, yielding

$$p_i^F = m/(k-1) \qquad \text{for } i \in I, \qquad p_i^F = 0 \text{ otherwise.}$$

Equation (7.11) yields

$$H(\mathbf{p}^F) = -V_{\text{tot}} + V(k-1) = -V_{\text{tot}} + Pm - V. \tag{7.60}$$

When $c_k^* > 0$, (7.59) and (7.60) must be compared to determine the optimum. If $\sum_1^k c_i^* \geq 1$, then $H(\mathbf{p}^E) \leq H(\mathbf{p}^F)$ and so $\mathbf{p}^F$ is optimal. If $\sum_1^k c_i^* \leq 1$, then $H(\mathbf{p}^E) \geq H(\mathbf{p}^F)$ and $\mathbf{p}^E$ is optimal. This concludes the discussion of Example 1′.

To complete this discussion of Model 3, we must consider the consequences of removing the restriction, introduced shortly below (7.45), that E is empty. Denote by P^* the problem of choosing an inspector's strategy $\mathbf{p}$ to minimize the expression (7.45). For each integer k in the range $0 \leq k \leq \min\{|E|, m\}$, let P_k be the problem obtained from P^* by adjoining the constraints

$$p_i = 1 \qquad \text{for exactly } k \text{ members } i \in E. \tag{7.61}$$

If $\mathbf{p}^k$ denotes an optimal solution to P_k, and K is such that

$$H(\mathbf{p}^K) = \max\{H(\mathbf{p}^k): \ 0 \leq k \leq \min(|E|, m)\}, \tag{7.62}$$

then clearly $\mathbf{p}^K$ is an optimal solution to P^*. Thus it suffices to be able to solve the problems P_k.

For each k, let E_k be a subset of E such that $|E| = k$ and $\sum\{c_i^*: \ I \in E_k\}$ is minimum. In other words, this sum consists of the k smallest members of $\{c_i^*: \ i \in E\}$; ties can be broken arbitrarily. It will be shown in section 8 that P_k has an optimal solution for which

$$p_i = 1 \qquad \text{for } i \in E_k, \tag{7.63}$$

$$p_i = \max\{0, m - k - (n - |E|)\}/(|E| - k) \qquad \text{for } i \in E - E_k. \tag{7.64}$$

Thus the analysis of P_k can be confined to such strategies. Because $m < n$, (7.64) gives $i \in L(\mathbf{p})$, and so (7.45) yields

$$\begin{aligned} -H(\mathbf{p}) = {} & \sum\{V_i c_i^*: \quad i \in E(p) \cap N_+ - E\} + V[L(\mathbf{p}) - E] \\ & + P[\sum\{c_i^*: \quad i \in E_k\} + |E - E_k|]. \end{aligned} \tag{7.65}$$

Therefore the problem P_k becomes that of choosing the nonnegative quantities $\{p_i: \quad i \in N - E\}$, which by (7.63) and (7.64) must sum to $\min\{m - k, n - |E|\}$, so as to minimize the sum of the first two summands in (7.65). This problem, however, is of the type treated earlier (no $V_i = P$), with N replaced by $N - E$ and m by $\min\{m - k, n - |E|\}$, and in that sense[4] can be regarded as "already solved."

It will also be shown in section 8 that the range of k, for problems P_k to be treated as indicated above, can be contracted. To define the reduced range, let $U = \{i: \quad V_i < P\}$ as before, and set

$$\rho = \min\{|E|, m - l - V(U)/P\},$$

$$\sigma = \max\{-1, m - V(U)/P\}.$$

Let r be the greatest integer not exceeding ρ, and s the smallest integer greater than σ. If $s \geq \min\{|E|, m\}$, set l equal to this minimum; otherwise, set l equal to the greatest integer between s and $\min\{|E|, m\}$ inclusive for which $\sum\{c_i^*: \quad i \in E_l - E_s\} < 1$, with tie-breakings in the choices of $E_l - E_s$ performed so that the latter is a subset of the former. Then the problems P_k need be solved only over the range

$$\max\{0, r\} \leq k \leq l. \tag{7.66}$$

No loss of alternative optima is incurred by imposing the upper limit; the same is true of the lower limit, when $\{c_i^*: \quad i \in E\}$ has as many as r members < 1.

EXAMPLE 1″: *Equal-Sized Firms*. We can now complete the discussion of this scenario, for Model 3 without assumption (7.9), by covering the case excluded in the previous Example 1′: that of all $V_i = P$. As before, we assume the numbering is such that

$$0 \leq c_1^* \leq c_2^* \leq \ldots \leq c_n^* \leq 1,$$

and can therefore take $E_k = \{1, 2, \ldots, k\}$. Since $E = N$, the first two terms on the right-hand side of (7.65) are zero, yielding

$$H(\mathbf{p}) = -P[\sum_1^k c_i^* + (n - k)]$$

as the optimal value (call it H_k) for problem P_k. Because

$$H_{k+1} - H_k = P(1 - c_{k+1}^*) \geq 0,$$

the value K in (7.62) can be taken here to be $K = m$. Thus an optimal strategy for the inspector is

[4] Except for the possibility of equality in $\min(m - k, n - |E|) = m' \leq n' = n - |E|$, when the solution of the reduced problem is trivial.

given, by (7.63) and (7.64), as

$$p_i = 1 \qquad \text{for } 1 \le i \le m,$$
$$p_i = 0 \qquad \text{for } m < i \le n,$$

and the optimum inspector's payoff is

$$H_m = -P[\textstyle\sum_1^m c_i^* + (n - m)] = -V_{\text{tot}} + Pm - V\sum_1^m c_i^*$$

consistent with the result (7.59) for Example 1′.

In order to illustrate the various types of optimal strategies which can occur, we will close this section with a numerical example.

EXAMPLE 3: Let $n = 5$, $m = 2$, $P = 1$, and

$$V_1 = V_2 = 1,\ V_3 = 0.3,\ V_4 = V_5 = 0.5.$$

Thus T is empty, $E = \{1,2\}$, and $V(U) = 1.3$. Also, $\rho = -0.3$ and $r = -1$, so that the lower limit in (7.66) does not ease the analysis. Without loss of generality, we can assume $c_1^* \le c_2^*$ and $c_4^* \le c_5^*$. We have $\sigma = 0.7$ and $s = 1$, so that if $c_2^* = 1$ then $l = 1$ and so (7.66) would excuse us from analyzing problem P_2. The problems P_k ($k = 0,1,2$) will be considered separately.

For P_2 we have $E_2 = \{1,2\}$. By (7.63),

$$p_1 = p_2 = 1,$$

and, in the reduction of P_2 to devices $\{D_3, D_4, D_5\}$, we have $n' = 3$ and $m' = 0$. Clearly this has the sole solution

$$p_3 = p_4 = p_5 = 0.$$

Combining these, we get the unique optimal solution to P_2. By (7.65),

$$-H_2 = 1.3 + c_1^* + c_2^*. \tag{7.67}$$

For P_1 we can take $E_1 = \{1\}$. By (7.63–7.64),

$$p_1 = 1, \qquad p_2 = 0,$$

and, in the reduction of P_1 to $\{D_3, D_4, D_5\}$ we have $n' = 3$ and $m' = 1$. According to (7.51) through (7.54), finding a best exceptional strategy $\mathbf{p}^E$ for this reduced problem is equivalent to finding $\{y_3, y_4, y_5\}$, each either 0 or 1, to

$$\text{minimize} \quad 0.3\, c_3^*\, y_3 + 0.5\, c_4^*\, y_4 + 0.5\, c_5^*\, y_5$$

subject to

$$0.3\, y_3 + 0.5\, y_4 + 0.5\, y_5 = 1.$$

The only solution of the constraint in (0,1)-valued variables is $y_3 = 0$, $y_4 = y_5 = 1$, yielding

$$p_3^E = 0, \quad p_4^E = p_5^E = 1/2,$$

and corresponding via (7.65) to an H-value given by

$$-H_1^F = 1.3 + c_1^* + 0.5\,c_4^* + 0.5\,c_5^*. \tag{7.68}$$

According to (7.21) through (7.23) finding a best violation-free strategy p^F for the reduced problem begins by finding $\{x_3, x_4, x_5\}$, all 0 or 1, to

$$\text{maximize} \quad 0.3\,x_3 + 0.5\,x_4 + 0.5\,x_5$$

subject to

$$0.3\,x_3 + 0.5\,x_4 + 0.5\,x_5 \le 1. \tag{7.69}$$

Since the maximum achieves equality in (7.69), if $c_5^* > 0$ then (7.13) is to be replaced by (7.13'), i.e., (7.69) is to be replaced by

$$0.3\,x_3 + 0.5\,x_4 + 0.5\,x_5 < 1.$$

This yields alternate solutions $I = \{3,4\}$ and $I = \{3,5\}$; the first of them gives, by (7.18–7.19),

$$p_3^F = 5/12, \quad p_4^F = 7/12, \quad p_5^F = 0.$$

By (7.65),

$$-H_1^F = 1.5 + c_1^*. \tag{7.70}$$

Comparison of (7.68) and (7.69) yields

$$H_1^F = \max\{H_1^E, H_1^F\} \qquad \text{if } c_4^* + c_5^* \le 0.41$$

$$H_1^F = \max\{H_1^E, H_1^F\} \qquad \text{if } c_4^* + c_5^* \ge 0.41.$$

Finally, corresponding to problem P_0 we have the empty set E_0. By (7.64),

$$p_1 = p_2 = 0.$$

The reduction of P_0 to $\{D_3, D_4, D_5\}$ has $n' = 3$ and $m' = 2$. Since no subset of $\{V_3, V_4, V_5\}$ sums to Pm', we need only find the best violation-free strategy for this reduced problem. This leads to $I = \{3, 4, 5\}$, and by (7.18–7.19),

$$p_3 = 10/17, \qquad p_4 = p_5 = 12/17.$$

By (7.65),

$$-H_0 = 2. \tag{7.71}$$

The maximum value $H_{\max}$ of H, corresponding to the optimal inspector's strategy, is given by

$$\begin{aligned} -H_{\max} &= \min\{-H_2, -H_1^E, -H_1^F, -H_0\} \\ &= \min\{1.3 + c_1^* + c_2^*,\ 1.3 + c_1^* + 0.5(c_4^* + c_5^*),\ 1.5 + c_1^*,\ 2\}. \end{aligned} \tag{7.72}$$

Note that the value of c_3^* is irrelevant. In table 3, it is shown that each of the four possible optimal strategies described above, and correspondingly each of the four candidate expressions for the minimum in (7.72), can yield the optimum for some choices of $\{c_1^*, c_2^*, c_4^*, c_5^*\}$. In each row of the table, the optimal value of H is marked with an asterisk.

TABLE 3. *Alternative forms of the optimum for Example 3*

	Problem P_2	Problem P_1		Problem P_0
	$p_1 = p_2 = 1$	$p_1 = 1$	$p_2 = 0$	$p_1 = p_2 = 0$
	$p_3 = p_4 = p_5 = 0$	$p_3^F = 0\ p_4^F = p_5^F = \frac{1}{2}$	$p_3^F = 5/12,\ p_4^F = 7/12,\ p_5^F = 0$	$p_3 = 10/17,$ $p_4 = p_5 = 12/17$
$c_1^* = c_2^* = 0.1,$ $c_4^* = c_5^* = 0.9$	-1.5^*	-2.3	-1.6	-2
$c_1^* = 0.1,\ c_2^* = 0.9,$ $c_4^* = c_5^* = 0.1$	-2.3	-1.5^*	-1.6	-2
$c_1^* = 0.1,\ c_2^* = 0.9,$ $c_4^* = c_5^* = 0.9$	-2.3	-2.3	-1.6^*	-2
$c_1^* = c_2^* = c_4^*$ $= c_5^* = 0.9$	-3.1	-3.1	-2.4	-2^*

8. Verifications for Model 3

In this section we present proofs of several assertions made during the analysis of Model 3 in section 7.

For the first of these, recall from (7.10) the definition

$$I(\mathbf{p}) = \{i: \quad p_i \geq V_i/P\}. \tag{8.1}$$

We are to show that a subset I of N has the form $I(\mathbf{p})$, for at least one $\mathbf{p}$, if and only if I satisfies the three conditions

$$V(I) \leq Pm, \tag{8.2}$$

$$I \subseteq \bar{T} = N - T, \tag{8.3}$$

$$P(m - |I| - |T|) < V(N - I - T). \tag{8.4}$$

For the proof of sufficiency, suppose I obeys (8.2–8.4). If $|I| + |T| < m$ then, following (7.16), set

$$p_i = 1, \qquad \text{for } i \in I \cup T, \tag{8.5}$$

$$p_i = [V_i(m - |I| - |T|)]/V(N - I - T) \qquad \text{for } i \in N - I - T \tag{8.6}$$

Since $m < n$ and all $V_i > 0$, the denominator in (8.6) is strictly positive. It is clear that all $p_i \geq 0$, and it follows from (8.4) and (8.6) that $p_i < 1$ for $i \in N - I - T$, so that all $p_i \leq 1$. From (8.6) we have

$$\sum \{p_i: \quad i \in N - I - T\} = m - |I| - |T|,$$

so that $\sum_1^n p_i = m$. Thus (8.5–8.6) indeed define a proper **p**. It follows from (8.3) and (8.5) that $i \in I$ implies i $\in I(\mathbf{p})$. It follows from (8.6) and (8.4) that $i \in N - I$ implies $i \in N - I(\mathbf{p})$. So $I = I(\mathbf{p})$, as desired.

If $|I| \leq m \leq |I| + |T|$, then following (7.17), set

$$p_i = 1 \qquad \text{for } i \in I, \tag{8.7}$$

$$p_i = [m - |I|]/|T| \qquad \text{for } i \in T, \tag{8.8}$$

$$p_i = 0 \qquad \text{for } i \in N - I - T. \tag{8.9}$$

The denominator of (8.8) can be nonpositive only when T is empty. Clearly $0 \leq p_i \leq 1$ for all i. Also, $\sum_1^n p_i = m$. Thus, (8.7–8.9) define a proper **p**. Clearly (8.7) implies that $i \in I(\mathbf{p})$ for all $i \in I$ and, if $i \in N - I$, then (8.8–8.9) imply that $i \in N - I(\mathbf{p})$. So $I = I(p)$, as desired.

If $|I| > m$ then, following (7.18–7.19), set

$$p_i = V_i/P + (1 - V_i/P)\,[Pm - V(I)]/[P|I| - V(I)] \qquad \text{for } i \in I, \tag{8.10}$$

$$p_i = 0 \qquad \text{for } i \in N - I. \tag{8.11}$$

In view of (8.2), the denominator of (8.10) could be nonpositive only if $P|I| \leq Pm$, contradicting $|I| > m$. It follows, using (8.10), (8.2), and (8.3), that all $p_i \geq 0$. Using $|I| > m$, it follows from (8.10) that $p_i \leq 1$ for $i \in I$, and thus for all i. Also

$$\begin{aligned}\textstyle\sum_1^n p_i &= V(I)/P + [|I| - V(I)/P][Pm - V(I)]/[P|I| - V(I)] \\ &= V(I)/P + (1/P)[Pm - V(I)] = m.\end{aligned}$$

Thus (8.10–8.11) define a proper **p**. Clearly (8.11) implies that $i \in N - I(\mathbf{p})$ for all $i \in N \geq I$. And if $i \in I$, then (8.10) together with the previous reasoning, implies that $p_i \geq V_i/P$, assuring $i \in I(\mathbf{p})$. Thus again $I = I(\mathbf{p})$, as desired.

For the necessity proof, we show that each set $I(\mathbf{p})$ satisfies (8.2–8.4). First,

$$\begin{aligned}V[I(\mathbf{p})] = \textstyle\sum \{V_i\colon\ i \in I(\mathbf{p})\} &\leq \sum \{Pp_i\colon\ i \in I(p)\} \\ &\leq P \textstyle\sum_1^n p_i = Pm,\end{aligned}$$

verifying (8.2). Second, by (8.1), $i \in I(\mathbf{p})$ implies $1 \geq V_i/P$, verifying (8.3). Third,

$$\begin{aligned}m = \textstyle\sum_1^n p_i &= \sum \{p_i\colon\ i \in \bar{T} - I(\mathbf{p})\} + \sum \{p_i\colon\ i \in I(\mathbf{p})\} + \sum \{p_i\colon\ i \in T\} \\ &\leq \sum \{V_i/P\colon\ i \in \bar{T} - I(\mathbf{p})\} + |I(\mathbf{p})| + |T| \\ &= \{V(\bar{T}) - V[I(\mathbf{p})]\}/P + |I(\mathbf{p})| + |T|,\end{aligned}$$

with strict inequality when $\bar{T} - I(\mathbf{p})$ is nonempty. This verifies (8.4) unless $I(\mathbf{p}) = T$ in which case (8.4) follows from $n > m$.

Next suppose **p** is an optimal strategy for the inspector, that $E = \{i\colon\ V_i = P\}$ is empty, and that

$$E(\mathbf{p}) \cap N_+ = \{i\colon\ p_i = V_i/P \text{ and } c_i^* > 0\} \tag{8.12}$$

is nonempty. We are to prove that

$$M(\mathbf{p}) = \{j\colon\ p_j > V_j/P\} \text{ is empty.} \tag{8.13}$$

and that

$$p_k < V_k/P \text{ implies } p_k = 0. \tag{8.14}$$

For the proof, first observe that $p_i < 1$ (i.e., $V_i < P$) for each $i \in E(\mathbf{p}) \cap N_+$. If there existed $j \in M(\mathbf{p})$, then one could choose a number δ with

$$0 < \delta < \min \{1 - V_i/P, p_j - V_j/P\}$$

and then define an inspector's strategy $\mathbf{q}$ by

$$q_i = p_i + \delta, \quad q_j = p_j - \delta, \quad q_t = p_t \text{ for } t \neq i,j.$$

Since $M(\mathbf{q}) = M(\mathbf{p}) \cup \{i\}$, $E(\mathbf{q}) = E(\mathbf{p}) - \{i\}$ and $L(\mathbf{q}) = L(\mathbf{p})$, it follows from (7.45) that

$$-H(\mathbf{q}) = -H(\mathbf{p}) - V_i c_i^* < -H(\mathbf{p}),$$

contradicting the optimality of $\mathbf{p}$. Hence (8.10) holds. If there existed an index k for which (8.14) was contradicted, i.e., $k \in L(\mathbf{p})$ and $p_k > 0$, then one could choose a number η with

$$0 < \eta < \min \{1 - V_i/P, p_k\}$$

and then define an inspector's strategy $\mathbf{r}$ by

$$r_i = p_i + \eta, \quad r_k = p_k - \eta, \quad r_t = p_t \text{ for } t \neq i, k.$$

As above, this would lead to a contradiction of the optimality of $\mathbf{p}$, so (8.14) must hold.

It remains to verify the assertions made in the final part of section 7, that following the discussion of Example 1′. Recall that P^* is the problem of choosing an inspector's strategy $\mathbf{p}$ to minimize

$$\begin{aligned} -H(\mathbf{p}) &= \sum \{V_i c_i^* : \ i \in E(\mathbf{p})\} + V[L(\mathbf{p})] \\ &= \sum \{V_i c_i^* : \ i \in E(\mathbf{p}) - E\} + V[L(\mathbf{p}) - E] \\ &+ P[\sum \{c_i^* : \ i \in E(\mathbf{p}) \cap E\} + |L(p) \cap E|]. \end{aligned} \tag{8.15}$$

For $0 \leq k \leq \min \{|E|, m\}$, P_k is the problem of minimizing H subject to the further restriction that $|E(\mathbf{p}) \cap E| = k$.

The first of these assertions is that P_k has an optimal solution in which $E(\mathbf{p}) \cap E = E_k$, where E_k is any particular solution to the problem of finding a subset E' of E which minimizes $\sum \{c_i^* : \ i \in E'\}$ subject to $|E'| = k$. In fact, the following argument will show that *every* optimal solution $\mathbf{p}$ of P_k is such that $E(\mathbf{p}) \cap E$ is a solution of the last-mentioned problem, which we denote Q_k.

For the proof, assume $\mathbf{p}$ is an optimal solution of P_k for which $E(\mathbf{p}) \cap E \neq E_k$. Since these two sets each have k members, it follows that $0 < k < |E|$ and that there exist indices

$$i \in E_k - E(\mathbf{p}) \cap E, \qquad j \in E(\mathbf{p}) \cap E - E_k.$$

By the definition of E_k, we have $c_i^* \leq c_j^*$, with strict inequality if $E(\mathbf{p}) \cap E$ is not a solution of Q_k (i.e., does not correspond to k smallest members of the indexed set $\{c_r^* : \ r \in E\}$). Define an inspector's strategy $\mathbf{q}$ by setting

$$q_j = p_i < 1, \qquad q_i = p_j = 1, \qquad q_t = p_t \qquad \text{for } t \neq i,j.$$

Then $E(\mathbf{q}) - E = E(\mathbf{p}) - E$, $L(\mathbf{q}) - E = L(\mathbf{p}) - E$, and

$$E(\mathbf{q}) \cap E = [E(\mathbf{p}) \cap E - \{j\}] \cup \{i\},$$

$$L(\mathbf{q}) \cap E = [L(\mathbf{p}) \cap E - \{i\}] \cup \{j\}.$$

Thus $\mathbf{q}$ satisfies the condition defining P_k, and it follows from (8.15) that

$$-H(\mathbf{p}) - [-H(\mathbf{q})] = P(c_j - c_i) \geq 0,$$

where the inequality is strict if $E(\mathbf{p}) \cap E$ is not a solution of Q_k, contradicting the optimality of $\mathbf{p}$ for P_k. Thus equality must hold, and so $\mathbf{q}$ is also optimal for P_k; repetition of this step will clearly lead to an optimal strategy π for P_k such that $E(\pi) \cap E = E_k$, as desired.

We are now assured that P_k has an optimal strategy $\mathbf{p}$ such that

$$p_i = 1 \qquad \text{for } i \in E_k, \tag{8.16}$$

$$p_i < 1 \qquad \text{for } i \in E - E_k. \tag{8.17}$$

By (8.15), for such a strategy $H(\mathbf{p})$ does not depend on the specific values of $\{p_i\colon\ i \in E - E_k\} = \{p_i\colon\ i \in L(\mathbf{p}) \cap E\}$. To see how these values might be chosen, note that by (8.16), the devices $\{D_i\colon\ i \in E_k\}$ receive an amount k from the total $\sum_1^n p_i = m$ of "inspection resources." For an optimal solution to P_k, since $H(\mathbf{p})$ does not depend on $\{p_i\colon\ i \in E - E_k\}$ so long as they are < 1, it can certainly do no harm to make available to the remaining $n - |E|$ devices $\{D_i\colon\ i \in N - E\}$ as much of the remaining subtotal $m - k$ of inspection resources as they can absorb. This amount

$$m - k - \min\{m - k, n - |E|\} = \max\{0, m - k - (n - |E|)\}$$

is to be realized as $\sum\{p_i\colon\ i \in E - E_k\}$, subject to (8.17). The simplest such allocation of this residual is the uniform one, leading to (7.63) and (7.64); alternative allocations lead to alternative optimal solutions to P_k.

For the verification of (7.66), first observe that problem P_k has been reduced to that of choosing the probabilities $\{p_i\colon\ i \in N - E = U \cup T\}$, summing to min $\{m - k, |U \cup T|\}$, so as to minimize the function given by (7.65), namely

$$\begin{aligned} -H(\mathbf{p}) = {} & \sum \{V_i c_i^*\colon\ i \in E(\mathbf{p}) - E\} + V[L(\mathbf{p}) - E] \\ & + P[\sum \{c_i^*\colon\ i \in E_k\} + |E - E_k|]. \end{aligned} \tag{8.18}$$

Let H_k denote the corresponding maximum value of H. The solution of the original problem, P^*, has been reduced to determining an index k for which max $\{H_k\colon\ 0 \leq k \leq m\}$ occurs, and solving the associated problem P_k.

The set-inclusions

$$E(\mathbf{p}) - E \subseteq U, \qquad T \subseteq L(\mathbf{p}) - E$$

are easily verified. Their use shows that (8.18) can be written

$$\begin{aligned} -H(\mathbf{p}) = {} & \sum \{V_i c_i^*\colon\ i \in E(\mathbf{p}) \cap U\} + V[L(\mathbf{p}) \cap U] \\ & + V(T) + P[\sum \{c_i^*\colon\ i \in E_k\} + |E - E_k|]. \end{aligned} \tag{8.19}$$

Thus solving P_k is equivalent to choosing the probabilities $\{p_i: \ i \in U\}$, summing to min $\{m - k, |U|\}$, so as to minimize the sum of the first two summands on the right-hand side in (8.19). The probabilities $\{p_i: \ i \in T\}$ can be chosen arbitrarily, except that they must sum to min $\{m - k, |U \cup T|\}$ − min $\{m - k, |U|\}$, which can be achieved since the difference lies between 0 and $|T|$ inclusive.

Let r be the largest integer not greater than

$$\rho = \min \{|E|, m - 1 - V(U)/P\}, \tag{8.20}$$

and assume for this and the next paragraph that $\rho > 0$, so that $r \geq 0$. Let $\mu = \min (|E|, m)$. We will next show that

$$\max \{H_k : r \leq k \leq \mu\} = \max \{H_k : 0 \leq k \leq \mu\}. \tag{8.21}$$

In other words, the search for the maximum in (7.62) can be confined to the subrange $\{k: \ r \leq k \leq \mu\}$ of $\{k: \ 0 \leq k \leq \mu\}$. Note that this is precisely the assertion made by the first equality in (7.66). The argument will also show that if r members of $\{c_i^*: \ \text{i} \in \text{E}\}$ are < 1, then no optima exist outside this subrange, i.e.,

$$\max \{H_k : 0 \leq k < r\} < \max \{H_k : r \leq k \leq \mu\}. \tag{8.22}$$

For the proof, assume $k \leq r$. Since $r \leq \rho$, it follows that $k < m - V(U)/P$. Moreover, since $V_i < P$ for $i \in U$, it follows that $V(U)/P < |U|$ if U is nonempty. Thus, if U is nonempty, then

$$V(U)/P < \min \{m - k, |U|\} \leq |U|.$$

This implies that the probabilities $\{p_i: \ i \in U\}$, summing to min $\{m - k, |U|\}$, can be chosen so that $p_i > V_i/P$. Such a choice (vacuously possible if U is empty) makes the first two summands on the right-hand side of (8.19) vanish, and therefore achieves the optimum in problem P_k. It follows that

$$H_k = - V(T) - P[S_k + |E - E_k|] \qquad \text{for } k \leq r, \tag{8.23}$$

where S_k denotes the sum of k smallest members of $\{c_i^*: \ \text{i} \in \text{E}\}$. Thus, for $k < r$,

$$H_{k+1} - H_k = P[1 - (S_{k+1} - S_k)] \geq 0,$$

where the inequality holds because $S_{k+1} - S_k$ is equal to some c_i^*. Moreover, if each summand of S_r is < 1, then for $k = r-1$ this inequality must be strict. The results (8.21) and (8.22) follow immediately from these observations.

Next let s be the smallest integer greater than

$$\sigma = \max \{-1, m - V(U)/P\}, \tag{8.24}$$

and consider any k with $k \leq \mu$ and

$$\sum \{c_i^*: \ i \in E_k - E_s\} \geq 1. \tag{8.25}$$

It will be shown below that $H_k < H_s$. This implies the second inequality of (7.66), completing the verification of the assertions in section 7. In (8.25), it is assumed that tie-breaking in the formation of E_k and E_s is performed so that the latter is a subset of the former; that is, the sum in (8.25) consists

of $k - s$ largest among k smallest members of $\{c_i^* : i \in E\}$, and is equal to $S_k - S_s$ in the notation of the preceding paragraph. If $s = 0$, E_s is the empty set.

The proof that $H_k < H_s$ will consist of showing that

$$-H_k \geq V(U) - P(m - k) + V(T) + P[S_k + |E| - k], \tag{8.26}$$

and that

$$-H_s < V(U) - P(m - s - 1) + V(T) + P[S_s + |E| - s]. \tag{8.27}$$

These inequalities together imply

$$\begin{aligned}(-H_s) - (-H_k) &< P(s + 1 - k) + P[S_s - S_k - (s - k)] \\ &= P[1 - (S_k - S_s)] \leq 0,\end{aligned}$$

where the last inequality follows from (8.25); this yields $H_s > H_k$.

To prove (8.26), let $\mathbf{p}^k$ be an optimal strategy for P_k, so that (8.19) yields

$$\begin{aligned}-H_k &\geq V[L(\mathbf{p}^k) \cap U] + V(T) + P[S_k + |E| - k] \\ &= V(U) - V[\{M(\mathbf{p}^k) \cup E(\mathbf{p}^k)\} \cap U] \\ &\quad + V(T) + P[S_k + |E| - k].\end{aligned} \tag{8.28}$$

Since

$$\begin{aligned}V[\{M(\mathbf{p}^k) \cup E(\mathbf{p}^k)\} \cap U] &\leq P \sum\{p_i^k: \quad i \in \{M(\mathbf{p}^k) \cap E(p^k)\} \cup U\} \\ &\leq P \sum\{p_i^k: \quad i \in U\} \leq P(m - k),\end{aligned}$$

(8.26) follows from (8.28).

We finally turn to the proof of (8.27). The definition of s [see (8.24)] implies $s \geq 0$. Since (8.25) implies $k > s$, we have $s < k \leq \mu$; thus $0 \leq s < \mu$, and so P_s and H_s are well-defined. Note that $s < \mu \leq m$; also, the definition of s gives $s > m - V(U)/P$. It follows that

$$0 < m - s < V(U)/P. \tag{8.29}$$

Since $V_i/P < 1$ for $i \in U$, (8.29) implies that there exists a subset W of U such that

$$m - s - 1 < V(W)/P < m - s. \tag{8.30}$$

Since integer $m - s > 0$, the first inequality implies that W is nonempty; since V_i/P for $i \in W$, it follows that

$$V(W)/P < \min\{m - s, |W|\}. \tag{8.31}$$

Thus there exist probabilities $\{p_i^*: \quad i \in W\}$, summing to the right-hand side of (8.31), for which $p_i^* > V_i/P$ for $i \in W$. Since

$$0 \leq \min\{m - s, |U|\} - \min\{m - s, |W|\} \leq |U - W|, \tag{8.32}$$

there exist probabilities $\{p_i^\dagger\colon\ i \in U - W\}$ summing to the middle term in (8.32). Thus the probabilities $\{p_i^\dagger\colon\ i \in U\}$ sum to min $\{m - s, |U|\}$, and so can be entered in (8.19) for the problem P_s, yielding

$$-H(\mathbf{p}^s) = \sum\{V_i c_i^*\colon\ i \in E(\mathbf{p}^s) \cap U\} + V[L(\mathbf{p}^s) \cap U] + V(T) + P[S_s + |E| - s].$$

Thus

$$\begin{aligned} -H_s \leq -H(\mathbf{p}^s) &\leq V[\{E(\mathbf{p}^s) \cup L(\mathbf{p}^s)\} \cap U] + V(T) + P[S_s + |E| - s] \\ &= V(U) - V[M(\mathbf{p}) \cap U] + V(T) + P[S_s + |E| - s]. \end{aligned}$$

Since the choice of $\{p_i^\dagger\colon\ i \in W\}$ assures $W \subseteq M(\mathbf{p}) \cap U$, it follows that

$$-H_s \leq V(U) - V(W) + V(T) + P[S_s + |E| - s],$$

and this together with the first inequality in (8.30) imply (8.27), completing the proof.

9. Index of Notation

	Page		Page
B_b	204	n_s	196
$\mathbf{c}$	191	N_m	202
c_i	191	N_m^+	202
$\mathbf{c}^\circ$	193	N_m^-	202
$c_i^\dagger$	194	N_m°	202
$\mathbf{c}(\mathbf{p})$	216	N_{m1}°	212
D_i	191	$\mathbf{p}$	191
E	194	P	191
EP	202	p_i	191
$E(\mathbf{p})$	217	$\mathbf{p}^\circ$	193
$F(\mathbf{c}, \mathbf{p})$	192	$p_i^\dagger$	194
F°	193	$\mathbf{p}^E$	226
$G(\mathbf{c}, \mathbf{p})$	202	$\mathbf{p}^F$	226
G°	202	T	193
$H(\mathbf{p})$	216	$\bar{T}$	193
$H(\mathbf{p}^E)$	225	U	194
$H(\mathbf{p}^F)$	225	V_i	191
$I(\mathbf{p})$	217	V_{tot}	195
L	204	V_b	196
$L(\mathbf{p})$	222	V_s	196
m	191	V_b^*	204
M	191		
$M(\mathbf{p})$	222	Δ	204
n	191	λ	205
N	191	α	204
n_b	196	β	204

(Paper 80B2–437)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

A Minimax-Measure Intersection Problem

Philip R. Meyers

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(March 16, 1976)

The problem solved is that of selecting n subsets of the unit interval, each of measure α, so as to minimize the maximum of the measures of their p-fold intersections. This is achieved by minimizing the sum of the measures of these p-fold intersections.

Key words: Combinatorial analysis; combinatorial probability; measure theory; minimax.

1. Introduction

Some years ago, NBS colleague S. Haber communicated the following problem: To select n subsets of the unit interval, each of measure $1/2$, so as to minimize the maximum of the measures of the pairwise intersections of these subsets. The problem is suggested by a paper [1][1] of Gillis which, settling "an unpublished conjecture of Erdos," proves that for denumerably infinite collections of sets of measure α, the value corresponding to the maximum pairwise-intersection measure has infimum α^2. (Collections with higher transfinite cardinality are treated by Gillis in [2].) Here we provide an explicit solution for collections of *finite* cardinalities n. Further, and also corresponding to [1], we consider as well the case of p-fold intersections with $2 \leq p \leq n$, and provide the corresponding explicit solution. (As noted in [2], the argument of [1] easily extends to show that α^p is the limiting value for a denumerably infinite collection.)

As preliminary, we introduce a second minimization and point out its relationship to our minimax problem, to wit: Select n subsets $A_1, A_2, \ldots, A_n$ of the unit interval, each of measure α, so that the *sum* of the measures of their p-fold intersections is minimum. If now $X = \{S_1, \ldots, S_n\}$, a solution to this minimum problem, can be chosen so that all its p-fold intersections have the *same* measure s, and if M is the maximum of the measures of the p-fold intersections of an arbitrary collection $A_1, A_2, \ldots, A_n$ with all $\mu(A_i) = \alpha$, then

$$\begin{aligned}\binom{n}{p} M &\geq \sum_{i_1<i_2<\ldots<i_p} \mu\left(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_p}\right) \\ &\geq \sum_{i_1<i_2<\ldots<i_p} \mu\left(S_{i_1} \cap S_{i_2} \cap \ldots \cap S_{i_p}\right) \\ &= \binom{n}{p} s.\end{aligned}$$

Thus $s \geq M$, demonstrating that X solves the minimax problem. This observation suggested the analysis which follows.

AMS Subject Classification: Primary 05A05; Secondary 05B99, 10E30, 28A75.

[1] Figures in brackets indicate the literature references at the end of this paper.

2. Analysis

We will use the following notation: $N = \{1,2,\ldots, n\}$; p is a fixed positive integer with $2 \leq p \leq n$. The underlying space is the unit interval I with Lebesgue measure μ (but the analysis actually carries over to any "atomless" probability space). Set complements of $A \subseteq I$ and $R \subseteq N$ are denoted A^c and R^c respectively. Let, for $0 \leq r \leq n$,

$$K_r = \{R \subseteq N: \quad |R| = r\}.$$

Given the real number α with $0 < \alpha \leq 1$, let

$$F(\alpha) = \{A \subseteq I : \mu(A) = \alpha\}$$

and let $F^n(\alpha)$ denote the n-fold Cartesian power of $F(\alpha)$, consisting of all n-tuples

$$X = \{A_1, A_2, \ldots, A_n\}$$

with each $A_i \in F(\alpha)$. For each such X, and each $R \subseteq N$, set

$$X_R = \{x \in I: \quad x \in A_i \qquad iff\ i \in R\};$$

an easily-proved property of these sets, to be used repeatedly below, is that

$$X_R \cap (\underset{i \in P}{\cap} A_i) = \begin{cases} X_R \text{ if } P \subseteq R, \\ \\ \phi \text{ otherwise.} \end{cases} \tag{1}$$

Note that the disjoint union

$$X_r = \underset{R \in K_r}{\cup} X_R$$

consists of those points $x \in I$ which are members of exactly r sets $A_i \in X$. Finally, for measurable B

$\subseteq I$, it is convenient to define

$$S(X,B) = \sum_{P \in K_p} \mu\left[\left(\cap_{i \in P} A_i\right) \cap B\right].$$

The "objective function" for the minimax problem is

$$M(X) = \max_{P \in K_p} \mu\left[\cap_{i \in P} A_i\right],$$

while that for the related minimization problem introduced in section 1 is

$$S(X) = S(X,I) = \sum_{P \in K_p} \mu\left[\cap_{i \in P} A_i\right].$$

An alternative formula for $S(X)$ will first be developed (Lemma 1), and then a necessary condition (Lemma 2) for some $X \in F^n(\alpha)$ to minimize S will be presented.

LEMMA 1: *For each* $X \in F^n(\alpha)$,

$$S(X) = \sum_{r=p}^{n} \binom{r}{p} \mu\left(X_r\right). \tag{2}$$

PROOF: Since $\{X_r\colon \ r = 0, 1, \ldots, n\}$ is a partition of I,

$$\begin{aligned} S(X) &= \sum_{r=0}^{n} S(X,X_r) \\ &= \sum_{r=0}^{n} \sum_{|R|=r} S(X,X_R). \end{aligned}$$

Applying (1) to each summand, we obtain

$$S(X) = \sum_{r=p}^{n} \sum_{|R|=r} \binom{r}{p} \mu\left(X_R\right),$$

yielding (2).

LEMMA 2. *If* X *minimizes* S *over* $F^n(\alpha)$, *and* $\mu(X_r) > 0$ *for some* $r \geq p$, *then* $\mu(X_t) = 0$ *for all* $t < r - 1$.

PROOF: Suppose, to the contrary, that there exist $r \geq p$ and $t < r - 1$ such that $\mu(X_r) > 0$ and $\mu(X_t) > 0$. We will prove the existence of an $X' \in F^n(\alpha)$ for which $S(X') < S(X)$, thus contradicting the hypothesis about X.

Since $\mu(X_r) > 0$ and $\mu(X_t) > 0$, K_r and K_t must contain respective members R and T with $\mu(X_R) > 0$ and $\mu(X_T) > 0$. Choose subsets Y and Z of I with

$$Y \subseteq X_R, \qquad Z \subseteq X_T, \qquad \mu(Y) = \mu(Z) > 0.$$

Also choose a member i of the nonempty set $R - T$; then

$$Y \subseteq A_i, \qquad Z \subseteq A_i^c.$$

Now define $X' = \{A_1, A_2, \ldots, A_i', \ldots, A_n\}$, where

$$A_i' = (A_i - Y) \cup Z;$$

Since $\mu(A_i') = \mu(A_i - Y) + \mu(Z) = \mu(A_i)$, we have $X' \in F^n(\alpha)$.

To prove that $S(X)') < S(X)$, observe that I is partitioned into Y, Z, and $I - Y - Z$.

Thus

$$S(X) = S(X,Y) + S(X,Z) + S(X,I - Y - Z),$$

$$S(X') = S(X',Y) + S(X',Z) + S(X',I - Y - Z).$$

Since X and X' differ only on $Y \cup Z$, it follows that

$$S(X) - S(X') = [S(X,Y) - S(X',Y)] - [S(X',Z) - S(X,Z)].$$

Since $Y \subseteq X_r$ and $Y \subseteq X'_{r-1}$, application of (1) to the summands of $S(X,Y)$ and $S(X',Y)$ yields

$$S(X,Y) - S(X',Y) = \binom{r}{p}\mu(Y) - \binom{r-1}{p}\mu(Y) = \binom{r-1}{p-1}\mu(Y).$$

Similarly, it follows from $Z \subseteq X_t$ and $Z \subseteq X'_{t+1}$ that

$$S(X',Z) - S(X,Z) = \binom{t+1}{p}\mu(Z) - \binom{t}{p}\mu(Z) = \binom{t}{p-1}\mu(Z).$$

Since $r - 1 > t$ and $\mu(Y) = \mu(Z) > 0$,

$$S(X) - S(X') = \binom{r-1}{p-1}\mu(Y) - \binom{t}{p-1}\mu(Z) > 0,$$

completing the proof.

We will subsequently show that if $\mu(X_r) > 0$ for some $r > p$ then $\mu(X_t) = 0$ for $t < r - 1$ is a sufficient condition for X to minimize S over $F^n(\alpha)$.

LEMMA 3: *For all* $X \in F^n(\alpha)$,

$$n\alpha = \sum_{r=0}^{n} \mu(X_r). \tag{3}$$

PROOF: Let c_i denote the characteristic function of A_i. Then

$$n\alpha = \sum_{i=1}^{n} \int_I c_i(x)\, d\mu(x) = \int_I [\textstyle\sum_{i=1}^{n} c_i(x)]\, d\mu(x)$$

$$= \sum_{r=0}^{n} \int_{X_r} [\textstyle\sum_{i=1}^{n} c_i(x)]\, d\mu(x) = \textstyle\sum_{r=0}^{n} r\mu(X_r).$$

It is now possible to prove:

LEMMA 4: If $n\alpha \leq p - 1$, then $S_{\min} = \min\{S(Y): \ Y \in F^n(\alpha)\} = 0$.

PROOF: It suffices to exhibit an $X \in F^n(\alpha)$ for which

$$\mu(X_r) = 0 \qquad \text{for } r \geq p. \tag{4}$$

To this end, let

$$A_i = [(i-1)\alpha, i\alpha) \pmod 1 \qquad \text{for } 1 \leq i \leq n.$$

Each point of $[0,1)$ corresponds (mod 1) to exactly $p - 1$ points of the interval $[0, p - 1)$, and thus to at most $p - 1$ points of the subinterval $[0, n\alpha)$; thus $X_r \cap [0, 1) = \phi$ for $r \geq p$, verifying (4).

LEMMA 5: *If for given* $X \in F^n(\alpha)$, *the largest* r *such that* $\mu(X_r) > 0$ *satisfies* $r \geq p$ *and further for* $t < r - 1$, $\mu(X_t) = 0$, *then*

$$S(X) = S_{\min}.$$

PROOF. It suffices to show that $S(X)$ has the same value for all $X \in F^n(\alpha)$ satisfying the conditions of the lemma. Consider such an X, and the greatest r for which $\mu(X_r) > 0$. Since

$$\sum_{t=0}^{n} \mu(X_t) = \mu(I) = 1, \tag{5}$$

such an r must exist. By the above condition $\mu(X_\theta) = 0$ for $t \neq r, r - 1$, and so by (5),

$$\mu(X_{r-1}) = 1 - \mu(X_r).$$

Let $n\alpha = m + \beta$ with m integral and $0 \leq \beta < 1$. It follows from (3) that

$$m + \beta = r\mu(X_r) + (r - 1)\,\mu(X_{r-1}) = (r - 1) + \mu(X_r), \tag{6}$$

and from (2) that

$$S(X) = \binom{r}{p}\mu(X_r) + \binom{r-1}{p}\mu\left(X_{r-1}\right) = \binom{r-1}{p} + \binom{r-1}{p-1}\mu(X_r). \tag{7}$$

If $\beta = 0$, then since m is integral and $0 < \mu(X_r) \leq 1$, it follows from (6) that $\mu(X_r) = 1$ and $r = m$, and then it follows from (7) that

$$S(X) = \binom{m-1}{p} + \binom{m-1}{p-1} = \binom{m}{p}. \tag{8}$$

If $\beta > 0$, then it follows from (6) that $\mu(X_r) = \beta$ and $m = r - 1$, and then it follows from (7) that

$$S(X) = \binom{m}{p} + \binom{m}{p-1}\beta. \tag{9}$$

Thus $S(X)$ is uniquely determined by the pair (m, β), i.e., by $n\alpha$. Note that (8) and (9) are consistent with Lemma 4, since both yield $S(X) = 0$ if $n\alpha \leq p - 1$.

We are now able to provide the solutions, both to the problem of minimizing $S(X)$ over $F^n(\alpha)$ and to the original problem of minimizing

$$M(X) = \max_{P \in K_p} \mu\left(\cap_{i \in P} A_i\right)$$

over $F^n(\alpha)$. Let $M_{\min}$ denote the value of this latter minimum. Then the solution takes the following form.

THEOREM. *Let* $n\alpha = m + \beta$ *with* m *integral and* $0 \leq \beta < 1$. *Then*

$$S_{\min} = M_{\min} = 0 \qquad \text{if } n\alpha \leq p - 1,$$

$$S_{\min} = \binom{m}{p} + \binom{m}{p-1}\beta, \qquad M_{\min} = S_{\min}/\binom{n}{p} \qquad \text{if } n\alpha > p - 1.$$

Thus, in particular, for the problem as originally posed where $p=2$ *and* $\alpha = 1/2$,

$$S_{\min} = \begin{cases} k(k-1)/2 & \text{if } n=2k \\ (k-1)^2/2 & \text{if } n=2k-1, \end{cases}$$

and $M_{\min} = (k - 1)/2(2k - 1)$.

PROOF: First suppose $n\alpha \leq p - 1$. Then $S_{\min} = 0$ follows from Lemma 4, whose proof constructed an $X \in F^n(\alpha)$ for which $\mu(\cup_{r=p}^{n} X_r) = 0$. Since every p-fold intersection of the members of X lies in this union, it follows that $M(X) = 0$, implying $M_{\min} = 0$.

Now suppose $n\alpha > p - 1$. The formula for $S_{\min}$ follows from (8) and (9). We will prove the result for $M_{\min}$ by constructing an $X \in F^n(\alpha)$ which satisfies the condition of Lemma 2, and which furthermore (see the end of sec. 1) has equal measures for each of its p-fold intersections. For this purpose, partition the interval $[0, \beta)$ into $\binom{n}{m+1}$ equal subintervals and the interval $[\beta, 1]$ into $\binom{n}{m}$ equal subintervals. Label the second family of subintervals as $\{X_M: \ M \in K_m\}$ and the first family as $\{X_Q: \ Q \in K_{m+1}\}$. Define

$$A_i = [\cup\{X_M: \ i \in M\}] \cup [\cup\{X_Q: \ i \in Q\}].$$

Then each A_i consists of $\binom{n-1}{m-1}$ intervals X_M and $\binom{n-1}{m}$ intervals X_Q, all disjoint, so all A_i have equal measure. If c_i denotes the characteristic function of A_i, then

$$\sum_{i=1}^{n} \mu(A_i) = \int_I \left(\sum_{i=1}^{n} c_i\right) d\mu = \int_0^\beta \left(\sum_{i=1}^{n} c_i\right) d\mu + \int_\beta^1 \left(\sum_{i=1}^{n} c_i\right) d\mu$$

$$= (m + 1)\beta + m(1 - \beta) = m + \beta = n\alpha.$$

Thus each $\mu(A_i) = \alpha$, i.e., $X \in F^n(\alpha)$. For $r \geq p$, $\mu(X_r) > 0$ holds only for $r = m$ and $r = m + 1$, so the condition of Lemma 2 is satisfied. The symmetry of the construction assures that all p-fold intersections of the members of X have equal measure; explicitly, for $P \in K_p$, we have

$$\bigcap_{i \in P} A_i = [\cup\{X_M: \ P \subset M\}] \cup [\cup\{X_Q: \ P \subset Q\}].$$

implying

$$\mu\left[\bigcap_{i \in P} A_i\right] = \binom{n-p}{m-p}(1 - \beta) / \binom{n}{m} + \binom{n-p}{m+1-p}\beta / \binom{n}{m+1},$$

independently of P.

3. References

[1] Gillis, J., Note on a property of measurable sets, J. London Math. Soc. **11**, 139–141 (1936).

[2] Gillis, J., Some combinatorial properties of measurable sets, Quart. J. Math. (Oxford) **7**, 191–198 (1936).

(Paper 80B2–438)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

Inverting Sparse Matrices by Tree Partitioning*

D. R. Shier

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(August 16, 1974)

This paper studies the tree partitioning of a graph whose definition is based on the pattern of zero elements present in a given matrix. This partitioning then indicates a particularly advantageous strategy for employing block Gaussian elimination over a certain class of matrices. The strategy is exploited for matrix inversion, where it is especially appropriate for problems which require finding only selected submatrices of the inverse. A graph-theoretic algorithm is given for automatically generating tree partitionings for any matrix. Combinatorial properties of this procedure are also discussed.

Key words: Block Gaussian elimination; graph; inversion; partitioning; sparse matrix; tree.

1. Introduction

The solution of "sparse" systems of linear equations has received considerable attention in recent years [13][1], [15], [19], and justifiably so, since coefficient matrices having relatively few nonzero entries arise quite frequently from physical problems [1], [4], [8]. The object of this paper is to discuss a partitioning method for inverting such sparse matrices which can rather readily adapt to the given zero-nonzero structure of a particular sparse matrix. This work is motivated by the observation [12] that the solution of a linear system is particularly simple when the graph underlying the coefficient matrix is exactly a tree. The present approach shows how a natural tree-like representation can be obtained for *any* matrix and how this leads to an especially simple method for *inverting* the given matrix. In addition, a straightforward extension of the method allows one to solve, rather effectively, systems of linear equations.

Henceforth, our attention will be focused on the class K of square matrices all of whose principal submatrices are nonsingular. Inasmuch as the class K contains all positive definite symmetric matrices (which do arise quite often in physically meaningful contexts), the method to be described will be applicable to a number of important situations. It is not necessary that the coefficient matrix A be symmetric; in fact, nonsymmetric matrices for which $A + A^T$ is positive definite also belong to K, as well as the nonsymmetric M-matrices [6] (having nonpositive off-diagonal elements and all principal minors positive).

2. Tree Partitionings

Let the real $n \times n$ matrix A be in K. Eventually, a simultaneous permutation of the rows and columns of A will be produced so that the resulting matrix has an especially simple form with identifiable (and exploitable) blocks of zero entries. To do this, we first associate a finite undirected *graph* G_A with the matrix $A = (a_{ij})$. The *nodes* of G_A correspond to the rows/columns of A, and an *edge* (i, j) will join node i to node j whenever $|a_{ij}| + |a_{ji}| > 0$.

In any graph G, the sequence of edges $(e_1, e_2, \ldots, e_k)$ forms a *chain* whenever $k + 1$ nodes $i_0, i_1, \ldots, i_k$ can be identified so that e_j joins nodes i_{j-1} and i_j for $j = 1, \ldots, k$; a chain is *elementary*

AMS Subject Classification: 05C05, 15A09.

*This work was done while the author was a National Research Council-National Academy of Sciences Postdoctoral Research Associate at the National Bureau of Standards, Washington, D.C. 20234.

[1] Figures in brackets indicate the literature references at the end of this paper.

if it does not meet the same node twice. A chain which consists of at least three distinct edges and which joins a node to itself is termed a *cycle*. If every pair of distinct nodes in a graph is joined by a chain, then the graph is said to be *connected*. A *tree* is a connected graph which contains no cycles.

It is now supposed that the nodes N of the graph G_A are partitioned into $m \geq 2$ distinct sets of nodes $N_1, N_2, \ldots, N_m$ which exhibit a tree structure when viewed as an undirected graph T_A. Namely, the nodes of T_A are the sets $N_1, N_2, \ldots, N_m$ and an edge joins N_i with N_j $(i = j)$ in T_A whenever an edge exists in G_A between some node in N_i and another node in N_j. The graph T_A so constructed is assumed then to be a tree. It will be shown in section 4 how to construct for connected graphs such a *tree partitioning*. (If the graph G_A is not connected then clearly A is decomposable as a block diagonal matrix; in this case the inverse of A is readily computable from the inverses of the diagonal blocks, each of which corresponds to a connected graph.) As an example, the "snowflake" graph [14, p. 209] depicted in figure 1 has the tree partitioning given in figure 2.

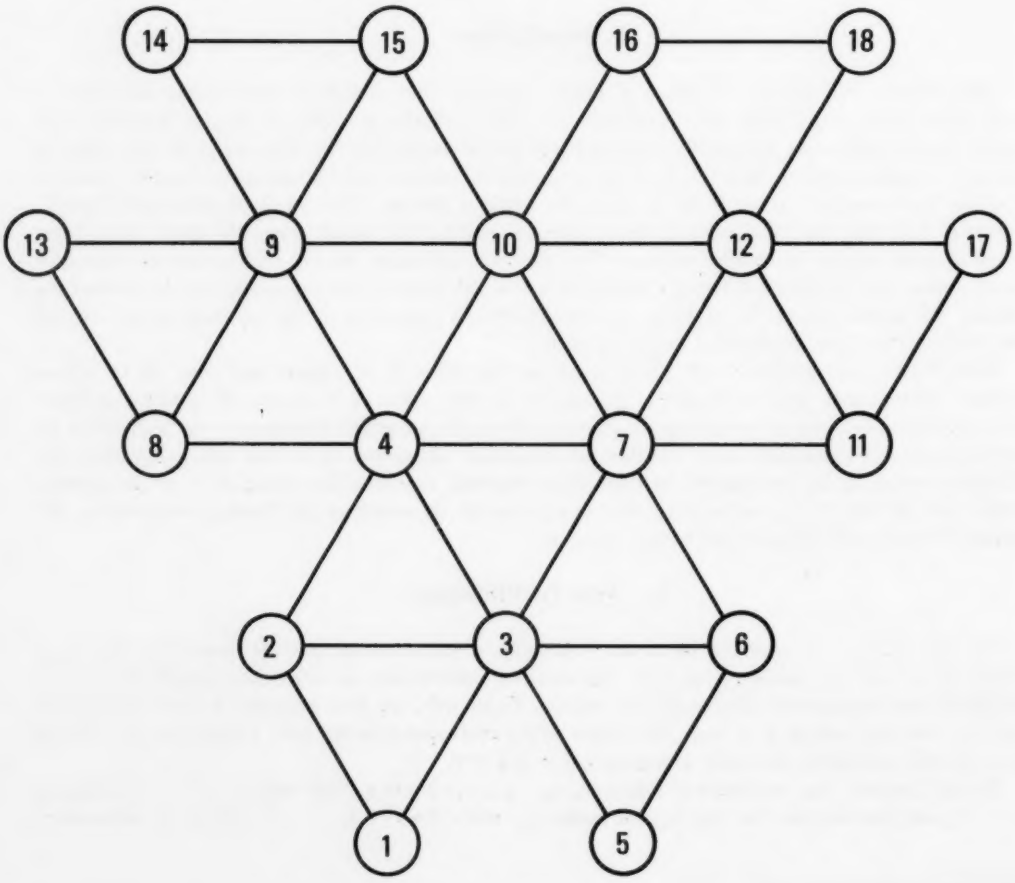


FIGURE 1. *An illustrative graph.*

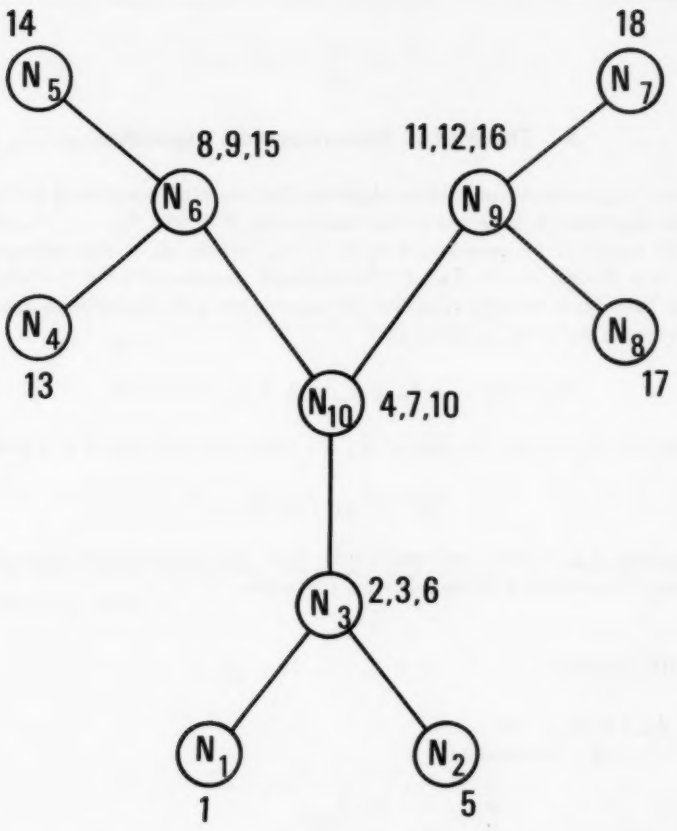


FIGURE 2. *A tree partitioning for the graph of figure 1 (the nodes comprising each node set are also shown).*

The concept of a tree partitioning has been previously applied [16] to shortest path calculations in large sparse networks. This concept can also be applied to the present situation of matrix inversion. The basic idea is that of decomposing the original problem into smaller subproblems, the solutions to which can be easily recombined to yield the information required from the original problem.

The present decomposition approach relies on a fundamental property of trees. This is, in each tree there always exists a *pendant* node—a node which is incident with precisely one edge. When such a pendant node i and its incident edge are deleted from a tree T, then a new tree $T - i$ is formed. This process of deleting a pendant node together with its incident edge can then be repeated anew using the tree $T - i$. Therefore, given the tree T_A, the node sets can be suitably relabelled so that N_1 is a pendant node of T_A, N_2 is a pendant node of $T_A - N_1$, . . ., and N_{m-1} is a pendant node of $T_A - N_1 - \ldots - N_{m-2}$. Such a labelling of T_A is shown in figure 2. Moreover, for each index $i \neq m$ there is associated a unique index $r(i)$ such that $N_{r(i)}$ is the next node set after N_i on the unique elementary chain joining N_i and N_m in the tree T_A. Because of the assumed labelling induced by the pendant node sets of T_A, it is always the case that $r(i) > i$. Table 1 displays these indices for the tree of figure 2.

TABLE 1. *The indices r(i) for the tree of figure 2.*

i	1	2	3	4	5	6	7	8	9
$r(i)$	3	3	10	6	6	10	9	9	10

3. The Matrix Decomposition Algorithm

In this section, a general decomposition algorithm for inverting matrices $A \in K$ will be presented and justified. This algorithm is based on a tree partitioning $N = (N_1, N_2, \ldots, N_m)$ of the associated graph G_A. Let the matrix A be partitioned as $A = (A_{ij})$, where A_{ij} is the submatrix composed of entries a_{uv} where $u \in N_i$ and $v \in N_j$. Let A^{-1} be similarly partitioned as $A^{-1} = (X_{ij})$. It is assumed that the node sets have been suitably relabelled in accordance with the scheme described in section 2. Hence, the ith row of A, $i \neq m$, is given by

$$A_i = (A_{i1}, \ldots, A_{ii}, 0, \ldots, 0, A_{i,r(i)}, 0, \ldots, 0), \tag{3.1}$$

where the submatrices $A_{ik} = 0$ $(k < i)$ unless $r(k) = i$. Also, the mth row of A is given by

$$A_m = (A_{m1}, \ldots, A_{mm}), \tag{3.2}$$

where the submatrices $A_{mk} = 0$ $(k < m)$ unless $r(k) = m$. The proposed decomposition algorithm for obtaining A^{-1} from A is embodied in the following procedure.

PROCEDURE INVERT.

1. Set $B_{ii} = A_{ii}$, $i = 1, \ldots, m$.
2. For $i = 1, \ldots, m - 1$ calculate

$$B_{i,r(i)} = -B_{ii}^{-1} A_{i,r(i)} \tag{3.3}$$

$$B_{r(i),i} = -A_{r(i),i} B_{ii}^{-1} \tag{3.4}$$

and perform the replacement $B_{r(i),r(i)} := B_{r(i),r(i)} + A_{r(i),i} B_{i,r(i)}$.

3. Compute $X_{mm} = B_{mm}^{-1}$ and for $i = m - 1, \ldots, 1$

$$X_{i,r(i)} = B_{i,r(i)} X_{r(i),r(i)}$$

$$X_{r(i),i} = X_{r(i),r(i)} B_{r(i),i}$$

$$X_{ii} = B_{ii}^{-1} + B_{i,r(i)} X_{r(i),i}.$$

4. The remaining X_{ij} blocks can be calculated using

$$X_{ij} = B_{i,r(i)} X_{r(i),j} \qquad i < j$$

$$X_{ij} = X_{i,r(j)} B_{r(j),j} \qquad i > j.$$

It is noted that this procedure requires the inversion of submatrices B_{ii} as well as the multiplication and addition of appropriate submatrices B_{jk}. Of course, there is no reason why the submatrices B_{ii} cannot be inverted (if still sufficiently sparse) by further applying the decomposition

algorithm to the subnetworks based on individual node sets N_i. Such an approach can be recursively applied until the sparsity of the B_{ii}'s has been exploited to its fullest.

The validity of the procedure will now be established. First, note that at the end of Step 2 of INVERT,

$$B_{ii} = A_{ii} + \sum_{r(k)=i} A_{ik} B_{ki}. \tag{3.5}$$

Since $A^{-1} = X = (X_{ij})$ is the solution to

$$AX = I, \tag{3.6}$$

then from (3.1) with $i \neq j$, $i < m$

$$\sum_{r(k)=i} A_{ik} X_{kj} + A_{ii} X_{ij} + A_{i,r(i)} X_{r(i)j} = 0. \tag{3.7}$$

It is now claimed that

$$X_{ij} = B_{i,r(i)} X_{r(i)j} \text{ if } i < j. \tag{3.8}$$

This assertion will be established inductively. Inasmuch as $i = r(k) > k$, equation (3.5) shows that $B_{11} = A_{11}$. Thus, using (3.7) with $j > 1$

$$A_{11} X_{1j} + A_{1,r(1)} X_{r(1)j} = 0,$$

whence from (3.3)

$$X_{1j} = -A_{11}^{-1} A_{1,r(1)} X_{r(1)j} = B_{1,r(1)} X_{r(1)j}.$$

Therefore, the assertion (3.8) holds for $i = 1$. Suppose the assertion is true for indices $k < i$, where $i < m$. Then from (3.7) with $i < j$

$$\sum_{r(k)=i} A_{ik} B_{ki} X_{ij} + A_{ii} X_{ij} + A_{i,r(i)} X_{r(i)j} = 0,$$

since $k < r(k) = i < j$ and so the inductive hypothesis can be applied. By (3.5) the above relation becomes

$$B_{ii} X_{ij} + A_{i,r(i)} X_{r(i)j} = 0,$$

whence

$$X_{ij} = -B_{ii}^{-1} A_{i,r(i)} X_{r(i)j} = B_{i,r(i)} X_{r(i)j},$$

using (3.3). Accordingly, the assertion (3.8) is established by induction.

In particular, choosing $j = r(i)$ in (3.8) yields

$$X_{i,r(i)} = B_{i,r(i)} X_{r(i),r(i)} \qquad \text{if } i < m. \tag{3.9}$$

Furthermore, from (3.2) and (3.6)

$$\sum_{r(k)=m} A_{mk}X_{km} + A_{mm}X_{mm} = I$$

or, applying (3.9),

$$\sum_{r(k)=m} A_{mk}B_{km}X_{mm} + A_{mm}X_{mm} = I,$$

so that from (3.5), $B_{mm}X_{mm} = I$, whereupon

$$X_{mm} = B_{mm}^{-1}. \tag{3.10}$$

Finally, from (3.1) and (3.6) with $i \neq m$

$$\sum_{r(k)=i} A_{ik}X_{ki} + A_{ii}X_{ii} + A_{i,r(i)}X_{r(i),i} = I$$

or, using (3.8) and (3.5),

$$B_{ii}X_{ii} + A_{i,r(i)}X_{r(i),i} = I.$$

By virtue of (3.3),

$$\begin{aligned} X_{ii} &= B_{ii}^{-1} - B_{ii}^{-1}A_{i,r(i)}X_{r(i),i} \\ &= B_{ii}^{-1} + B_{i,r(i)}X_{r(i),i}. \end{aligned} \tag{3.11}$$

The counterparts to equations (3.8) and (3.9) can be established by using the relation $X\,A = I$ instead of (3.6), thus yielding

$$X_{ij} = X_{i,r(i)}B_{r(i),j} \qquad \text{if } i > j \tag{3.12}$$

$$X_{r(i),i} = X_{r(i),r(i)}B_{r(i),i} \qquad \text{if } i < m. \tag{3.13}$$

Together, eqs (3.8) — (3.13) serve to establish the validity of procedure INVERT, provided that the submatrices B_{ii} encountered for inversion are nonsingular. This property of the B_{ii}'s will next be shown to follow from the fact that $A \in \mathrm{K}$. First, given any tree partitioning of G_A, the matrix $A = (A_{ij}) \in \mathrm{K}$ evidently satisfies the following condition:

Block Principal Submatrix (BPS) Condition. Given the matrix A partitioned into m^2 blocks by corresponding sets $N_1, N_2, \ldots, N_m$ of row and column indices, then any submatrix determined by deleting corresponding sets N_j (possibly none at all) is nonsingular.

It will now be shown that, under the transformation described by Step 2 of INVERT, this property is inherited by certain submatrices of the transformed matrix. Let $A = (A_{ij})$ be written as

$$A = M_0' = \left[\begin{array}{c|ccccc} B_{11} & 0 & \cdots & A_{1,r(1)} & \cdots & 0 \\ \hline 0 & & & & & \\ \vdots & & & & & \\ A_{r(1),1} & & & M_1 & & \\ \vdots & & & & & \\ 0 & & & & & \end{array}\right].$$

After the transformation with $i = 1$ in Step 2 of INVERT, the matrix M_1 will be transformed into a new matrix M_1' which differs from M_1 only in that the diagonal block $B_{r(1),r(1)}$ has been replaced by

$$\begin{aligned} B'_{r(1),r(1)} &= B_{r(1),r(1)} + A_{r(1),1} B_{1,r(1)} \\ &= B_{r(1),r(1)} - A_{r(1),1} B_{11}^{-1} A_{1,r(1)}. \end{aligned} \tag{3.14}$$

Now it is easily verified that $M_1' = (M_0'/B_{11})$, the *Schur complement* [2] of B_{11} in M_0', defined by:

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \Rightarrow (M/A) = D - CA^{-1}B.$$

Given that $A = M_0'$ is *BPS*, it is claimed that M_1' is also *BPS*. In fact, consider any principal submatrix S' determined by the blocks of M_1'. If the diagonal block $B'_{r(1),r(1)}$ does not occur in S', then S' is also a principal block submatrix of M_0', and hence is nonsingular. If S' does contain $B'_{r(1),r(1)}$, then consider the corresponding submatrix S of M_0' obtained by replacing $B'_{r(1),r(1)}$ by its pre-transformation value $B_{r(1),r(1)}$. Let T be the submatrix of M_0' defined by

$$T = \left[\begin{array}{c|ccccc} B_{11} & 0 & \cdots & A_{1,r(1)} & \cdots & 0 \\ \hline 0 & & & & & \\ \vdots & & & & & \\ A_{r(1),1} & & & S & & \\ \vdots & & & & & \\ 0 & & & & & \end{array}\right].$$

Since T and B_{11} are principal block submatrices of M_0' (which is *BPS*), then T and B_{11} are nonsingular. Moreover, by the determinantal formula of Schur [7]

$$\det T = \det B_{11} \cdot \det (T/B_{11}),$$

it follows that (T/B_{11}) is nonsingular as well. However, it is easily seen from (3.14) that $(T/B_{11}) = S'$ and so S' is nonsingular. Thus, M_1' is BPS.

More generally, the above argument can be repeated to show that if M_{k-1}' is BPS then the submatrix M_k defined by

$$M_{k-1}' = \left[\begin{array}{c|ccccc} B_{kk} & 0 & \cdots & A_{k,r(k)} & \cdots & 0 \\ \hline 0 & & & & & \\ \vdots & & & & & \\ A_{r(k),k} & & & M_k & & \\ \vdots & & & & & \\ 0 & & & & & \end{array}\right]$$

yields a transformed M_k' (after executing Step 2 of INVERT with $i = k$) which is also *BPS*. Given, then, that each M_k' is BPS we can conclude that in particular the diagonal block $B_{k+1,k+1}$ is nonsingular. This is precisely what is needed to ensure that at each iteration of Step 2, the matrix B_{ii}^{-1} is defined. Accordingly, the procedure INVERT is guaranteed to be well-defined and will then produce the matrix A^{-1} as required. It should be noted that if A satisfies the BPS condition for the given partition, then the INVERT procedure will remain valid, even if A is not a member of K.

In effect, procedure INVERT is a statement of block Gaussian elimination, which becomes especially simple and manageable in the context of a tree partitioning. It should be noted that in

Steps 2 and 3 the blocks of the inverse matrix (corresponding to nonzero blocks of the original matrix A) can be calculated without introducing any new "fill-ins"; this is just to say that the given zero blocks are preserved. Moreover, we see that the existence of a tree partitioning allows one to solve linear equations in very much the same way as inverting matrices. Rather than needing to perform submatrix inversions, we need instead to solve a linear subsystem of equations at each step.

4. Finding Tree Partitionings

In this section, an algorithm will be described for obtaining tree partitions of an undirected graph $G = (N,E)$. Without loss of generality, it may be assumed that G is a connected graph having at least two nodes. Indeed, any graph which is not connected can be resolved into at least two maximally connected subgraphs[2] or *connected components*. Clearly, then, the tree partitioning problem can be studied and solved separately with regard to each of these *connected* components.

The basic idea of the algorithm can be most easily explained by means of an example. Consider, therefore, the connected graph of figure 1 for which $N = \{1, \ldots, 18\}$. The process is begun by selecting a subset N_1 of nodes from $C_1 = N$ with $\phi \subset N_1 \subset C_1$;[3] say, $N_1 = \{4, 7, 10\}$. It is then easy to compute

$$\Gamma(N_1) = \{j \in C_1 - N_1: \quad (i, j) \in E \text{ for some } i \in N_1\},$$

the set of all nodes in C_1 adjacent with (but not including) nodes in N_1. For this example, $\Gamma(N_1) = \{2, 3, 6, 8, 9, 11, 12, 15, 16\}$. It is also easy to find the connected components of the subgraph based on the nodes $C_1 - N_1$ and thus the node sets C_k constituting each of these connected components. Here, $C_2 = \{8, 9, 13, 14, 15\}$, $C_3 = \{11, 12, 16, 17, 18\}$ and $C_4 = \{1, 2, 3, 5, 6\}$. The node sets N_2, N_3, and N_4 are then chosen as $N_2 = C_2 \cap \Gamma(N_1) = \{8, 9, 15\}$, $N_3 = C_3 \cap \Gamma(N_1) = \{11, 12, 16\}$, and $N_4 = C_4 \cap \Gamma(N_1) = \{2, 3, 6\}$.

The next iteration of the process computes

$$\begin{aligned}\Gamma(N_2) &= \{j \in C_2 - N_2: \quad (i, j) \in E \text{ for some } i \in N_2\} \\ &= \{13, 14\}\end{aligned}$$

and determines the sets $C_5 = \{13\}$ and $C_6 = \{14\}$ associated with the two connected components for $C_2 - N_2$. The node sets N_5 and N_6 are given by $N_5 = C_5 \cap \Gamma(N_2) = \{13\}$, and $N_6 = C_6 \cap \Gamma(N_2) = \{14\}$. In similar fashion, subsequent iterations produce $N_7 = \{17\}$, $N_8 = \{18\}$, $N_9 = \{1\}$, $N_{10} = \{5\}$. The tree partitioning which reflects this disposition of node sets (apart from relabelling) is shown in figure 2.

A general statement of the tree partitioning procedure for connected graphs $G = (N,E)$ is provided by

PROCEDURE TREEPART.

1. Let N_1 satisfy $\phi \subset N_1 \subset N$. Set $C_1 = N$, $p = 1$, $q = 1$.
2. Compute $\Gamma(N_p) = \{j \in C_p - N_p: \quad (i, j) \in E \text{ for some } i \in N_p\}$. If $\Gamma(N_p) = \phi$ then go to Step 3. Otherwise,
   a. Find the node sets $C_{q+1}, \ldots, C_{q+r}$ constituting the connected components of the subgraph determined by $C_p - N_p$.
   b. Form $N_{q+i} = C_{q+i} \cap \Gamma(N_p)$ for $i = 1, \ldots, r$.
   c. Join node set N_p by an edge to each of the node sets $N_{q+1}, \ldots, N_{q+r}$.
   d. Let $q := q + r$.
3. Let $p := p + 1$. If $p \leq q$ then go to Step 2. Otherwise, terminate.

In the above algorithm, the actual edges of the tree formed from the node sets N_k are generated at

[2] A *subgraph* of a given undirected graph $G = (N,E)$ has for its nodes a subset $N_0 \subseteq N$ and contains those edges of E which join nodes in N_0.
[3] The notation $A \subset B$ means that A is a proper subset of B.

Step 2c. Because N_1 is a nonempty proper subset of N, the resulting tree partitioning has always at least two and at most $n = |N|$ constituent node sets. Generally speaking, the less the density of edges in G, the greater will be the number of node sets in a tree partition for G. (This statement will be given a more precise form in the theorem which appears later in this section.)

The fact that G is connected ensures that each node of G will appear in some node set N_k. Moreover, the connectivity of G is also enough to guarantee that each of the node sets generated in Step 2b is nonempty. Indeed, suppose that $N_p \neq \phi$ and, say, $N_{q+1} = \phi$. This means that no edges exist between nodes in C_{q+1} and nodes in N_p. Since the nodes of C_{q+1} form a connected component relative to the nodes of $C_p - N_p$, then it follows that C_{q+1} forms a connected component of C_p. Because $N_p \neq \phi$, then $C_{q+1} \subset C_p$ and thus one connected component properly contains another. This is a manifest contradiction, and so $N_p \neq \phi$ must imply that $N_{q+1} \neq \phi, \ldots, N_{q+r} \neq \phi$. In view of the fact that $N_1 \neq \phi$, this observation can be repeatedly applied to demonstrate that all node sets N_k are nonempty. Since the $N_{q+1}, \ldots, N_{q+r}$ defined in Step 2b are nonempty, the edges generated in Step 2c correctly reflect the adjacency relations between the node sets in G. In addition, the fact that no two distinct node sets N_{q+i}, N_{q+j} formed in Step 2b are connected by a chain in $C_p - N_p$ (else C_{q+i} and C_{q+j} would be connected in $C_p - N_p$) is enough to guarantee that the procedure always generates a tree graph for the node sets N_k. Accordingly, it has been verified that TREEPART does indeed produce a tree partitioning for every (nontrivial) connected graph G.

It is interesting to note that when applied to a graph which is itself a tree this procedure yields the original tree back again whenever N_1 is chosen to consist of a single node. In general, different choices for the initial node set N_1 will lead to different tree partitionings. Thus, TREEPART allows one to obtain a whole range of different tree partition structures for the same graph G. In order to obtain a "good" decomposition of the original network (e.g., one containing a large number of node sets), a reasonable heuristic might be to choose an N_1 which consists of relatively few nodes and which itself forms a highly connected subgraph. It often seems desirable to choose N_1 so that $N - N_1$ has several connected components.

As noted above, the number m of node sets in a tree partitioning is at least 2 and at most n. A more precise upper bound on the number of node sets is given by the following result.

THEOREM: *If* G = (N,E) *where* |N| = n, |E| = k *then the number* m *of node sets in a tree partitioning of* G *satisfies*

$$m \leq \left[\frac{3 + \sqrt{4n^2 - 4n - 8k + 1}}{2} \right],$$

where [ ] *is the greatest integer function. Moreover, this upper bound is actually tight; that is, there exist graphs with* n *nodes and* k *edges for which equality obtains.*

First, it will be convenient to prove the following lemma[4].

LEMMA: *For a given tree structure* T *with node sets* $N_1, N_2, \ldots, N_m$, *the maximum number of edges possible in any underlying graph* G *is achieved when* $|N_i| = 1$ *for* $i \neq p$ *and* $|N_p| = n - m + 1$, *where* N_p *is a node set of maximal degree in* T.

PROOF: Given the tree structure T, regard the quantities $n_i = |N_i|$ as integer variables constrained by

$$n_i \geq 1, \quad \sum_{i=1}^{m} n_i = n. \tag{4.1}$$

[4] A referee has pointed out a simplified proof of the theorem which does not use this lemma. However, because the lemma refers to a more general problem and is of independent interest, we include its statement and proof.

For given n_i's, the maximum number of edges in the underlying graph G occurs when each N_i corresponds to a complete subgraph, and when for each $(N_i, N_j) \in T$ all of the possible $n_i n_j$ edges between the corresponding subgraphs of G are actually present in G. This maximum number is given by

$$F(\mathbf{n}) = F(n_1, \ldots, n_m) = \frac{1}{2}\sum_{i=1}^{m} n_i(n_i - 1) + \sum \{n_i n_j : \ (N_i, N_j) \in T\}.$$

We therefore address the problem of maximizing $F(\mathbf{n})$ subject to (4.1). Let $\mathbf{n}$ be an arbitrary assignment of weights to the nodes satisfying (4.1). It will be shown that there exists an assignment to the nodes of weights $\mathbf{n}^0$ having the form

$$\begin{aligned} n_i &= 1, \ i \neq p \\ n_p &= n - m + 1, \end{aligned} \tag{4.2}$$

with N_p a node set in T of maximal degree, such that $F(\mathbf{n}) \leq F(\mathbf{n}^0)$.

In order to show this, consider the following procedure which performs a reassignment for some current assignment $\mathbf{n}$ relative to the edge $(N_i, N_j) \in T$.

PROCEDURE REASSIGN (i, j).

1. Let w_i be the sum of weights n_s for all node sets N_s adjacent to node set N_i $(s \neq j)$ and let w_j be the sum of weights n_s for all node sets N_s adjacent to node set N_j $(s \neq i)$. Define $r = n_i + n_j$.
2. If $w_i > w_j$ then set $n_i' = r - 1$, $n_j' = 1$. Otherwise, set $n_i' = 1$, $n_j' = r - 1$.

It is clear that for this new assignment $\mathbf{n}'$, where only n_i and n_j have been changed, equation (4.1) is satisfied. Moreover, it will now be shown that $F(\mathbf{n}) \leq F(\mathbf{n}')$. Indeed, if $I = n_i$ and $J = n_j$ then

$$F(\mathbf{n}) = \alpha + \frac{1}{2}\{I(I - 1) + J(J - 1)\} + IJ + Iw_i + Jw_j,$$

where α is independent of I and J. Or, setting $I + J = r$,

$$F(\mathbf{n}) = \alpha + \frac{1}{2}r(r - 1) + Iw_i + (r - I)w_j = \beta + I(w_i - w_j),$$

where β is independent of I, J. Thus, if $w_i > w_j$ it is clearly advantageous, in the sense of increasing $F(\mathbf{n})$, to make I as large as is feasible; namely, $I = r - 1$. Similarly, if $w_i \leq w_j$ then setting $J = r - 1$ will increase $F(\mathbf{n})$ as much as possible. Since this is precisely what is being done in REASSIGN, one has $F(\mathbf{n}) \leq F(\mathbf{n}')$.

Suppose that procedure REASSIGN is performed so long as there are adjacent nodes N_i and N_j for which $n_i > 1$, $n_j > 1$. Consider then the set P of node sets N_i for which $n_i > 1$. Note that the lemma certainly holds when $P = \phi$. If $|P| \geq 2$, it will be demonstrated that a further reassignment of node weights n_i can be made which does not decrease $F(\mathbf{n})$ but which will reduce $|P|$.

Assume, then, that N_i and N_j are in P. Thus $(N_i, N_j) \notin T$ since otherwise procedure REASSIGN could have been employed. In fact, each node set N_k adjacent with N_i must have weight $n_k = 1$, and similarly for N_j. Let d_i be the degree of node set N_i, and d_j the degree of N_j. Also, set $I = n_i$, $J = n_j$ and $r = I + J$. It is claimed that $F(\mathbf{n})$ cannot be decreased by performing the reassignment

$$\begin{aligned} n_i &= 1, \ n_j = r - 1 \quad &&\text{if } d_j \geq d_i \\ n_j &= 1, \ n_i = r - 1 \quad &&\text{if } d_j < d_i. \end{aligned} \tag{4.3}$$

Indeed,

$$F(\mathbf{n}) = \alpha + \frac{1}{2}r(r-1) - IJ + Id_i + Jd_j$$

$$= \alpha + \frac{1}{2}r(r-1) - I(r-I) + Id_i + (r-I)d_j$$

$$= \beta + I^2 + (d_i - d_j - r)I,$$

where α, β are independent of I, J. Because $g(I) = I^2 + (d_i - d_j - r)I$ is a convex function of I, it is maximized over the interval $[1, r-1]$ at one of the endpoints. It is easy to verify that

$$g(1) \geq g(r-1) \quad \Leftrightarrow \quad (d_j - d_i)(r-2) \geq 0 \quad \Leftrightarrow \quad d_j \geq d_i,$$

since $r = I + J > 2$. Therefore, the reassignment given in (4.3) cannot decrease the value of $F(\mathbf{n})$. In other words, given node sets N_i, $N_j \in P$, one of the two sets can be made to have weight 1 without decreasing F, and so can be removed from P. By continuing this procedure, P eventually reduces to a single node set with weight $w > 1$. Finally, by using REASSIGN and (4.3) as necessary with respect to the node set N_p of weight $w > 1$ and each node set of weight 1, it will be assured that the resulting N_p is in fact a node of maximal degree d_p in T. Thus, the assignment which is ultimately produced is of the form (4.2). Since the value $F(\mathbf{n})$ is never decreased throughout the process, we indeed have $F(\mathbf{n}) \leq F(\mathbf{n}^0)$. Since the original assignment $\mathbf{n}$ was arbitrary and since all assignments $\mathbf{n}^0$ in (4.2) have the same value $F(\mathbf{n}^0)$, the lemma is proved.

By virtue of this lemma, the number k of edges in an undirected graph G with a tree partitioning of n nodes into m node sets satisfies

$$k \leq F(\mathbf{n}^0) = \frac{1}{2}(n-m+1)(n-m) + d_p(n-m+1) + (m-1-d_p),$$

since a tree on q nodes has precisely $q-1$ edges. Thus,

$$k \leq \frac{1}{2}(n-m+1)(n-m) + d_p(n-m) + (m-1)$$

$$= \alpha(n, m) + d_p(n-m),$$

with $\alpha(n, m)$ independent of d_p. Accordingly, the quantity $\{\alpha(n, m) + d_p(n-m)\}$ is maximized by choosing d_p as large as possible for fixed n, m: namely, $d_p = m-1$ (all other $d_i = 1$). In any event, then,

$$k \leq \frac{1}{2}(n-m+1)(n-m) + (m-1)(n-m+1)$$

$$= \frac{1}{2}(n^2 - n - 2 + 3m - m^2),$$

whence

$$h(m) \equiv m^2 - 3m - (n^2 - n - 2k - 2) \leq 0.$$

Now the roots m_1 and m_2 of $h(m) = 0$ are $m_1 = \frac{1}{2}(3 - \sqrt{D})$ and $m_2 = \frac{1}{2}(3 + \sqrt{D})$, where $D = 9 + 4(n^2 - n - 2k - 2)$. Because $k \leq \frac{1}{2}n(n - 1)$ then $D \geq 1$ whence $m_1 \leq 1$ and $m_2 \geq 2$. It follows that $h(m) \leq 0$ for all m with $1 \leq m \leq m_2$. Inasmuch as m must be a positive integer, then

$$m \leq [m_2] = \left[\frac{3 + \sqrt{4n^2 - 4n - 8k + 1}}{2}\right].$$

Therefore, the first part of the theorem is established. To prove the second half, consider the graph $\hat{G} = \hat{G}(n, m)$ which consists of $n - m + 1$ "central" nodes, every two of which are joined by an edge, and $m - 1$ "satellite" nodes, each of which is joined by an edge to every central node. The number of edges in $\hat{G}$ is thus

$$k = \frac{1}{2}(n - m + 1)(n - m) + (m - 1)(n - m + 1)$$

and the resulting $m_2 = \frac{1}{2}(3 + \sqrt{(2m - 3)^2}) = m$. Moreover, this graph on n nodes does admit of a tree partitioning with m node sets: namely, choose $m - 1$ node sets each containing a single satellite node, and an mth node set containing all central nodes. Accordingly, the second part of the theorem is verified.

It is worthwhile to note that when the underlying graph is a tree ($k = n - 1$), the upper bound provided by the theorem is exactly n. Here again the upper bound is tight since a tree on n nodes admits of a tree partitioning into n node sets (just let each N_i contain a single node). When the underlying graph is *complete* (that is, every pair of distinct nodes is joined by an edge) then a tree partitioning can have at most two node sets. In fact, the upper bound provided in the theorem for the situation $k = \frac{1}{2}n(n - 1)$ is seen to be 2 also. As a final illustration, the upper bound on m is calculated to be 17 for the graph depicted in figure 1 ($n = 18$, $k = 33$); the tree partitioning of this graph shown in figure 2 contains 10 node sets.

5. Computational Remarks

The principal virtue of the tree decomposition algorithm given in section 3 is that computations need only be performed on arrays which are significantly smaller than the original matrix. Accordingly, much larger matrices than could normally be accommodated in core can be inverted. Moreover, the form of partitioning employed seems to readily adapt itself to the particular zero-nonzero structure of the matrix being studied. Of course, the tree partitioning concept is even more appropriate when repeated matrix inversions are to be made for a sequence of matrices differing only in that nonzero entries are reestimated or varied parametrically [4], [9]. Indeed, for a fixed structure of sparseness the underlying graph remains the same and so the tree partitioning can be found once and for all. In addition, it should be pointed out that TREEPART is at worst an $0(mk)$ algorithm since no edge of the graph need be scanned more than m times. In practice, the amount of computational labor required to find a tree partitioning is really quite modest. For example, when the underlying graph has $k = 0(n)$, as would often be the case for sparse matrices (e.g., resulting from the rectangular and triangular lattices which arise in numerical solution of partial differential equations), the theorem of section 4 shows that $mk \leq 0(n^2)$; accordingly, TREEPART is *at worst* an $0(n^2)$ algorithm for these sparse graphs.

Moreover, the tree decomposition approach is able to exploit effectively the sparsity of the original matrix and thus reduce the amount of computation required for matrix inversion. Suppose,

for example, that each of the node sets N_i has $|N_i| = r$. Then the use of INVERT to find the inverse of A requires $0(m^2r^3) = 0(n^2 \cdot \frac{n}{m})$ operations; thus for a given number of nodes, the computational effort decreases as the number of node sets increases. Without exploiting sparsity, standard methods for finding A^{-1} necessitate $0(m^3r^3)$ operations. Furthermore, the decomposition procedure allows the user selectivity in calculating the submatrices X_{ij} of A^{-1}. Indeed, INVERT only *requires* the calculation of the $3m - 2$ submatrices X_{ij} which correspond to nonzero A_{ij} in the original matrix A. The remaining $(m - 1)(m - 2)$ submatrices can be calculated, *if desired*, during Step 4 of the algorithm. Thus, the present decomposition approach would be especially appropriate for problems which require finding only certain submatrices of A^{-1}; such a situation arises when one is interested in finding the variances of estimated coefficients in multiple linear regression, since the required variances are derived from the diagonal entries of a matrix $(X^TX)^{-1}$.

Several procedures have been described for transforming a given matrix into one with a particular partitioned block structure [3], [10], [11], [17], [18, Ch. 3]. The use of a tree partitioning seems sufficiently flexible to deal with a wide range of possible partitions. Fiedler [5] discusses a type of partitioning more general than that given here; however, the solution method indicated in [5] appears to involve too many submatrix inversions to be practically advantageous. Another approach to inverting matrices using graph-theoretic concepts has been described by Harary [10]. Such a method unfortunately is of little use when the matrix A is irreducible. A generalization of Harary's method is presented in [3]. In addition, Mayoh [11] and Steward [17] have discussed techniques for permuting the rows and columns of A so that the resulting matrix has a particularly simple form. The computational requirements of such techniques are, however, difficult to assess.

6. References

[1] Churchill, M. E. A sparse matrix procedure for power systems analysis programs, in Large Sparse Sets of Linear Equations, J. K. Reid, Ed., (Academic Press, London, 1971), pp. 127–138.

[2] Cottle, R. W. Manifestations of the Schur complement, Linear Algebra and Appl. **8** (1974), pp. 189–211.

[3] Dulmage, A. L. and Mendelsohn, N. S. On the inversion of sparse matrices, Math. Comp. **16** (1962), pp. 494–496.

[4] Erisman, A. M. Decomposition methods using sparse matrix techniques with application to certain electrical network problems, in Decomposition of Large-Scale Problems, D. M. Himmelblau, Ed., (North-Holland Publishing Co., Amsterdam, 1973), pp. 69–80.

[5] Fiedler, M. On inverting partitioned matrices, Czechoslovak Math. J. **13** (1963), pp. 574–586.

[6] Fiedler, M. and Pták, V., On matrices with non-positive off-diagonal elements and positive principal minors, Czechoslovak Math. J. **12** (1962), pp. 382–400.

[7] Gantmacher, F. R., The Theory of Matrices, Vol. I, (Chelsea Publishing Co., New York, 1960).

[8] George, A., Nested dissection of a regular finite element mesh, SIAM J. Numer. Anal. **10** (1973), pp. 345–363.

[9] Gustavson, F. G., Liniger, W. and Willoughby, R., Symbolic generation of an optimal Crout algorithm for sparse systems of linear equations, J. Assoc. Comput. Mach. **17** (1970), pp. 87–109.

[10] Harary, F., A graph theoretic approach to matrix inversion by partitioning, Numer. Math. **4** (1962), pp. 128–135.

[11] Mayoh, B. H., A graph technique for inverting certain matrices, Math. Comp. **19** (1965), pp. 644–646.

[12] Parter, S., The use of linear graphs in Gauss elimination, SIAM Rev. **3** (1961), pp. 119–130.

[13] Reid, J. K., Ed., Large Sparse Sets of Linear Equations, (Academic Press, London, 1971).

[14] Rose, D. J., A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations, in Graph Theory and Computing, R. C. Read, Ed., (Academic Press, New York, 1972), pp. 183–217.

[15] Rose, D. J. and Willoughby, R. A., Eds., Sparse Matrices and Their Applications, (Plenum Press, New York, 1972).

[16] Shier, D. R., A decomposition algorithm for optimality problems in tree-structured networks, Discrete Math. **6** (1973), pp. 175–189.

[17] Steward, D. V., Partitioning and tearing systems of equations, J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. **2** (1965), pp. 345–365.

[18] Tewarson, R. P., Sparse Matrices, (Academic Press, New York, 1973).

[19] Willoughby, R. A., Ed., Sparse Matrix Proceedings, IBM Research Report RA1, Yorktown Heights, New York, 1969.

(Paper 80B2–439)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

The Structure of Higher Degree Symmetry Classes of Tensors*

Russell Merris**

(March 22, 1976)

The paper is concerned with symmetry classes of tensors which arise from a permutation group G and irreducible character χ of G. In case χ is of degree 1, a well-known algorithm is available for inducing a basis of the symmetry class from the underlying vector space. When the degree of χ is greater than 1, no comparable construction has been discovered. The difficulties are discussed and results obtained in some special cases.

Key words: Decomposable (or pure) tensor products; irreducible complex character; orthogonality relations; permutation group.

1. Introduction

Let V be a complex inner product space of dimension n. Let $\overset{m}{\otimes}V$ denote the mth tensor power of V, and let $v_1\otimes \ldots \otimes v_m$ be the (pure or decomposable) tensor product of the indicated vectors. The inner product in V induces an inner product in $\overset{m}{\otimes}V$ which is completely determined by its action on the set of decomposable tensors, namely

$$(v_1\otimes \ldots \otimes v_m, w_1\otimes \ldots \otimes w_m) = \prod_{t=1}^{m} (v_t, w_t). \tag{1}$$

By S_m, we mean the full symmetric permutation group on $\{1, \ldots, m\}$. If $\sigma \in S_m$, there is a (unique) linear operator $P(\sigma^{-1})$ on $\overset{m}{\otimes}V$ which has the effect $P(\sigma^{-1})v_1\otimes \ldots \otimes v_m = v_{\sigma(1)}\otimes \ldots \otimes v_{\sigma(m)}$, for all $v_1, \ldots, v_m \in V$. It follows that $P(\sigma)\,P(\pi) = P(\sigma\pi)$. Moreover, from (1), $P(\sigma)^* = P(\sigma^{-1})$. Let G be a subgroup of S_m, and χ an irreducible (complex) character of G. Define

$$T(G, \chi) = \frac{\chi(id)}{o(G)} \sum_{\sigma\in G} \chi(\sigma)\, P(\sigma),$$

where id = identity of G, and $o(G)$ is the order of G. By the orthogonality relations for characters, $T(G, \chi)$ is an orthogonal projection onto its range $V_\chi(G)$ (see, e.g., [5][1] or [12]). The subspace $V_\chi(G)$ is called a symmetry class of tensors [8]. Several authors have exploited these symmetry classes to obtain information about so called generalized matrix functions (see, e.g., [5], [8], [9], and [11]).

Until recently, however, most of the work has involved only linear characters. One reason for this preference is the existence, in the case $\chi(id) = 1$, of a convenient basis for $V_\chi(G)$ which is induced from a given basis of V. In the case $\chi(id) > 1$, it is not easy to obtain such a basis. A more precise idea of our interest must await further introductory material.

With $\Gamma_{m,n}$ we denote the set of functions from the first m positive integers to the first n. It is convenient to think of $\Gamma_{m,n}$ as a set of integer sequences of length m. Thus

AMS Subject Classification: Primary, 15A69; Secondary, 20C15.
*An invited paper. The research leading to this article was supported in part by the National Science Foundation.
**Present address: Department of Mathematics, California State University, Hayward, CA 94542.

$$\Gamma_{m,n} = \{\gamma = (\gamma(1), \ldots, \gamma(m)): \quad 1 \le \gamma(t) \le n, \qquad 1 \le t \le m\}.$$

If $e_1, \ldots, e_n$ is an orthonormal basis of V, it is well known (see, e.g., [8]) that $\{e_\gamma^\otimes = e_{\gamma(1)} \otimes \ldots \otimes e_{\gamma(m)} : \gamma \in \Gamma_{m,n}\}$ is an o.n. basis of $\otimes^m V$. It follows that $\{e_\gamma^* = T(G, \chi)e_\gamma^\otimes : \gamma \in \Gamma_{m,n}\}$ must span $V_\chi(G)$. (In general, write $x_1 * \ldots * v_m = T(G, \chi)v_1 \otimes \ldots \otimes v_m$.) If $\alpha, \beta \in \Gamma_{m,n}$, observe that

$$\begin{aligned}(e_\alpha^*, e_\beta^*) &= (T(G,\chi)e_\alpha^\otimes, T(G,\chi)e_\beta^\otimes) \\ &= (T(G,\chi)e_\alpha^\otimes, e_\beta^\otimes) \\ &= \frac{\chi(id)}{o(G)} \sum_{\sigma \in G} \chi(\sigma^{-1}) \prod_{t=1}^{m} (e_{\alpha\sigma(t)}, e_{\beta(t)}).\end{aligned} \tag{2}$$

It follows from (2) that $(e_\alpha^*, e_\beta^*) = 0$ unless there is a $\pi \in G$ such that $\beta = \alpha\pi$. We will say that $\alpha \equiv \beta \pmod G$ if there exists a $\pi \in G$ such that $\beta = \alpha\pi$. Clearly, "$\equiv \pmod G$" is an equivalence relation.

If $\beta = \alpha\pi$, for some fixed $\pi \in G$, then

$$\begin{aligned}(e_\alpha^*, e_\beta^*) &= \frac{\chi(id)}{o(G)} \sum_{\sigma \in G} \chi(\sigma) \prod_{t=1}^{m} (e_{\alpha(t)}, e_{\alpha\pi\sigma(t)}) \\ &= \frac{\chi(id)}{o(G)} \sum_{\tau \in G} \chi(\pi^{-1}\tau) \prod_{t=1}^{m} (e_{\alpha(t)}, e_{\alpha\tau(t)}) \\ &= \frac{\chi(id)}{o(G)} \sum_{\tau \in G_\alpha} \chi(\pi^{-1}\tau),\end{aligned} \tag{3}$$

where $G_\alpha = \{\tau \in G: \quad \alpha\tau = \alpha\}$ is the stabilizer subgroup of α. In particular, by taking $\pi = id$ in (3) one sees that $e_\alpha^* \neq 0$, if and only if

$$\alpha \in \Omega = \{\gamma \in \Gamma_{m,n}: \quad \sum_{\sigma \in G_\gamma} \chi(\sigma) \neq 0\},$$

i.e., Ω consists of those sequences γ which have the property that the restriction of χ to G_γ contains the identically 1 character as a component. (Although not explicit in the notation, Ω depends on m, n, G and χ.) It follows that $\{e_\omega^*: \omega \in \Omega\}$ spans $V_\chi(G)$.

Now, if $\alpha \equiv \beta \pmod G$, then G_α is conjugate to G_β. Therefore, Ω is a union of equivalence classes, i.e., if $\alpha \equiv \beta \pmod G$, then $e_\alpha^* = 0$ if and only if $e_\beta^* = 0$. Let $\bar{\Delta}$ be a system of distinct representatives for the equivalence classes in Ω. (In practice, $\bar{\Delta}$ is usually chosen to consist of those elements of Ω which come first, in lexicographic order, in their equivalence classes.) Then

$$\Omega = \bigcup_{\alpha \in \bar{\Delta}} \{\alpha\sigma: \quad \sigma \in G\}. \tag{4}$$

THEOREM A ([10]): *Let* $e_1, \ldots, e_n$ *be a basis of* V. *Then* $V_\chi(G)$ *is the direct sum of the spaces* $\langle e_{\alpha\sigma}^*: \quad \sigma \in G\rangle$, *as* α *ranges over* $\bar{\Delta}$. (*The angular brackets denote linear closure.*)

PROOF: Choose the inner product on V with respect to which $e_1, \ldots, e_n$ is orthonormal. The theorem follows from (4) and the definitions.

The result which makes the degree one case so fruitful is this:

THEOREM B (Marcus and Minc [9]): *Let* $e_1, \ldots, e_n$ *be a basis of* V. *Suppose* $\chi(\mathrm{id}) = 1$. *Then* $\{e_\alpha^*: \quad \alpha \in \bar{\Delta}\}$ *is a basis of* $V_\chi(G)$.

PROOF: It is routine to verify that $P(\sigma)$ commutes with $T(G, \chi)$ for all $\sigma \in G$. Moreover, if $\chi(id) = 1$, then $P(\sigma)T(G, \chi) = \chi(\sigma^{-1})T(G, \chi)$. It follows that $e^*_{\alpha\sigma} = \chi(\sigma)e^*_\alpha$ if $\chi(id) = 1$. So, each subspace in the direct sum of Theorem A is one dimensional.

That $\{e^*_\alpha: \alpha \in \bar{\Delta}\}$ is not a basis of $V_\chi(G)$ when $\chi(id) > 1$ is evident from the following result of S. Pierce [12]:

THEOREM C: *Let $\alpha \in \bar{\Delta}$ be arbitrary. There is a $\sigma \in G$ such that e^*_α and $e^*_{\alpha\sigma}$ are linearly independent if and only if $\chi(\mathrm{id}) > 1$.*

R. Freese [5] has improved Theorem C. Let $s_\alpha = \dim\langle e^*_{\alpha\sigma}: \sigma \in G\rangle$. Freese's result is this:

THEOREM D. *If $\alpha \in \Gamma_{m,n}$, then*

$$s_\alpha = \chi(\mathrm{id})\,(\chi, 1)_{G_\alpha},$$

i.e., s_α is $\chi(\mathrm{id})$ times the number of occurrences of the identically one character in the restriction of χ to G_α.

To conclude this section, we list a number of facts about s_α which follow from our discussion above.

(i) $s_\alpha \neq 0$, if and only if $\alpha \in \Omega$.

(ii) $\sum_{\alpha \in \bar{\Delta}} s_\alpha = \dim V_\chi(G)$

(iii) $\chi(id) \leq s_\alpha < \chi(id)^2$, for all $\alpha \in \Omega$.

(iv) $s_\alpha \leq [G:G_\alpha]$, for all α. (In fact, it is clear from Freese's proof of Theorem D that $s_\alpha < [G:G_\alpha]$ unless χ is identically 1 and $G = G_\alpha$.)

(v) $\|e^*_\alpha\|^2 = s_\alpha/[G:G_\alpha]$, if $e_1, \ldots, e_n$ is an o.n. basis of V. (See (3).)

2. Results

Presently, the outstanding problem is to choose from $\{e^*_{\alpha\sigma}: \sigma \in G\}$ a basis of $\langle e^*_{\alpha\sigma}: \sigma \in G\rangle$. In this generality, the task seems quite difficult. We are able to supply an answer (Theorem 4 below) only in a very special situation.

As a first step toward analyzing the dependence relations among the elements of $\{e^*_{\alpha\sigma}: \sigma \in G\}$, $\alpha \in \Omega$, one is naturally led to consider

$$G^\alpha = \{\sigma \in G: \text{ there exists } c_\alpha(\sigma) \text{ such that } e^*_{\alpha\sigma} = c_\alpha(\sigma)e^*_\alpha\}.$$

(If $\chi(id) = 1$, then $G^\alpha = G$ and $c_\alpha = \chi$. Moreover, $G_\alpha \subseteq G^\alpha$ for all $\alpha \in \Omega$.)

We first claim that G^α does not depend on the basis $e_1, \ldots, e_n$. Let $v_1, \ldots, v_n$ be another basis of V. Define a linear operator T on V by $T(e_i) = v_i$, $1 \leq i \leq n$, and linear extension. It is well known (see, e.g., [10]) that T induces a linear operator $K(T)$ on $V_\chi(G)$ such that

$$K(T)(x_1 * \ldots * x_m) = (Tx_1) * \ldots * (Tx_m),$$

for all $x_1, \ldots, x_m \in V$. Since T is invertible, it follows that $K(T)$ is invertible. Indeed, $K(T)^{-1} = K(T^{-1})$. Applying $K(T)$ to both sides of the equation $e^*_{\alpha\sigma} = c_\alpha(\sigma)\, e^*_\alpha$ one obtains $v^*_{\alpha\sigma} = c_\alpha(\sigma)\, v^*_\alpha$.

THEOREM 1: *For all $\alpha \in \Omega$, G^α is a group and c_α is a linear character on it.*

PROOF: If $\sigma, \pi \in G$, then

$$e^*_{\alpha\,\sigma\,\pi} = P(\pi^{-1})\, e^*_{\alpha\sigma}$$

$$= c_\alpha(\sigma)\, P(\pi^{-1})\, e^*_\alpha$$

$$= c_\alpha(\sigma)\, e^*_{\alpha\pi}$$

$$= c_\alpha(\sigma)\, c_\alpha(\pi)\, e^*_\alpha.$$

Thus $\sigma\ \pi \in G^\alpha$ and, since $e^*_\alpha \neq 0, c_\alpha(\sigma\pi) = c_\alpha(\sigma)\, c_\alpha(\pi)$.

We remark (without proof since the result seems peripheral to the present undertaking) that the restriction of χ to G^α contains c_α as a component for all $\alpha \in \Omega$, i.e.,

$$(\chi, 1)_{G_\alpha} > 0 \text{ implies } (\chi, c_\alpha)_{G^\alpha} > 0.$$

The converse fails.

COROLLARY 1: *If* $\alpha \in \Omega$, *then* $s_\alpha \leq [G\colon\ G^\alpha]$. (*Indeed, if* S^α *is a system of right coset representatives for* G^α *in* G, *then* $\bigcup_{\alpha \in \Delta} \{e^*_{\alpha\pi}\colon\ \pi \in S^\alpha\}$ *spans* $V_\chi(G)$.)

(It follows from Corollary 1 and (iii) of section 1 that $\chi(id) \leq [G\colon\ G^\alpha]$ for all $\alpha \in \Omega$. This inequality may be of some interest in itself because G^α is generally not normal in G [1, Theorem (53.17)].)

EXAMPLE 1: Let $G = S_3$. Let χ be the irreducible character of G of degree 2, and take $\alpha = (1, 1, 2)$. Then $G_\alpha = S_2$. If G_α were not all of G^α, then G^α would be all of S_3, implying that $[G\colon\ G^\alpha] = 1 < s_\alpha = 2$, contradicting Corollary 1. Therefore, $G_\alpha = G^\alpha$, and $[G\colon\ G^\alpha] = 3$. In particular, it's not true in general that $s_\alpha = [G\colon\ G^\alpha]$.

Subsequent developments will make clearer the relationship between G_α and G^α. We now make another definition. Let G be a subgroup of S_m. Let χ be an irreducible character of G. Define

$$G_\chi = \{\sigma \in G\colon\ |\chi(\sigma)| = \chi(id)\}.$$

It is easy to see that G_χ is a normal subgroup of G and $\lambda = \chi/\chi(id)$ is a linear character on it [4, p. 35], [11]. In fact, G_χ consists of those σ which are represented by scalars in any representation which affords χ.

THEOREM 2: *For* $\alpha \in \Omega$, $G_\chi \subseteq G^\alpha$, *and the restriction of* c_α to G_χ is λ.

PROOF: Let $\sigma \in G_\chi$. Then

$$e^*_{\alpha\sigma} = \frac{\chi(id)}{o(G)} \sum_{\pi \in G} \chi(\pi) P(\pi\sigma^{-1})\, e^*_\alpha$$

$$= \frac{\chi(id)}{o(G)} \sum_{\pi \in G} \chi(\pi\sigma) P(\pi)\, e^*_\alpha$$

$$= \frac{\chi(id)}{o(G)}\, \lambda(\sigma) \sum_{\pi \in G} \chi(\pi) P(\pi)\, e^*_\alpha$$

$$= \lambda(\sigma)\, e^*_\alpha.$$

COROLLARY 2: *For all* $\alpha \in \Omega$, $G_\alpha G_\chi \subseteq G^\alpha$.

EXAMPLE 2: It is tempting to conjecture that $G_\alpha G_\chi = G^\alpha$. Unfortunately, this is not always the case. Let $G = S_5$. Suppose χ arises from the frame (3, 2). Let $\alpha = (1,1,1,2,3)$ and let σ be the transposition (45). Then $\alpha\sigma = (1,1,1,3,2)$, and a brute force computation shows that $e^*_\alpha = e^*_{\alpha\sigma}$. In particular, since $G\chi = \{id\}$ and $G_\alpha = S_3$, it follows that $(45) \in G^\alpha \setminus G_\alpha G_\chi$.

It was proved in [11] that $\chi(id)^2 \leq [G: G_\chi]$, so the inequality $s_\alpha \leq [G: G_\chi]$ which arises from Theorem 2 is not very interesting. However, one might be tempted to conjecture that $\chi(id)^2 \leq [G: G_\alpha G_\chi]$ for all $\alpha \in \Omega$. A counterexample follows.

EXAMPLE 3. Let G be the subgroup of S_4 generated by $\{(14)(23), (1234)\}$. Then G is the dihedral group D_4 of order 8. Let χ be the irreducible character of G of degree 2. Then $\chi(id) = 2 = -\chi((13)(24))$, and χ is zero on the rest of G. Thus, $G_\chi = \{id, (13)(24)\}$. If $\alpha = (1,1,2,2)$, then $G_\alpha = \{id, (12)(34)\}$, and $\alpha \in \bar{\Delta}$. Moreover, $G_\alpha G_\chi = \{id, (12)(34), (13)(24), (14)(23)\}$ and $[G: G_\alpha G_\chi] = 2$, which is less than $\chi(id)^2 = 4$.

It is worth pointing out some other features of Example 3: Since

$$1 < s_\alpha \leq [G: G^\alpha] \leq [G: G_\alpha G_\chi] = 2,$$

it follows that $s_\alpha = [G: G_\alpha G_\chi]$, and hence $G^\alpha = G_\alpha G_\chi$. Moreover, $\chi(id)^2 = [G: G_\chi]$. In a moment, we shall see that these observations are connected. First, however, it should be mentioned that the case of equality in $\chi(id)^2 \leq [G: G_\chi]$ is related to some recent work of F. DeMeyer, S. M. Gagola, G. Janusz, K. M. Timmer, and J. Yellen ([2], [3], [6], [13], and [14]) in which the case of equality in $\chi(id)^2 \leq [G: Z(G)]$ is studied. In particular, since $Z(G) \subseteq G_\chi$, $[G: Z(G)] = \chi(id)^2$ implies $[G: G_\chi] = \chi(id)^2$.

THEOREM 3: *If* $\chi(id)^2 = [G: G_\chi]$, *then* $s_\alpha = [G: G_\alpha G_\chi]$ (*and therefore* $G^\alpha = G_\alpha G_\chi$) *for all* $\alpha \in \Omega$.

PROOF: If $[G: G_\chi] = \chi(id)^2$, then $\chi(\sigma) \neq 0$, if and only if $\sigma \in G_\chi$ [11]. It follows from Theorem D that

$$s_\alpha = \frac{\chi(id)}{o(G_\alpha)} \sum_{\sigma \in G_\alpha} \chi(\sigma)$$

$$= \frac{\chi(id)^2}{o(G_\alpha)} \sum_{\sigma \in G_\alpha \cap G_\chi} \lambda(\sigma),$$

where $\lambda = \chi/\chi(id)$. Since $s_\alpha \neq 0$, and since λ is a linear character, it must be that $\lambda(\sigma) = 1$ for all $\sigma \in G_\alpha \cap G_\chi$. Thus

$$s_\alpha = \chi(id)^2\, o(G_\alpha \cap G_\chi)/o(G_\alpha)$$

$$= o(G)\, o(G_\alpha \cap G_\chi)/o(G_\chi)\, o(G_\alpha)$$

$$= [G: G_\alpha G_\chi]$$

from elementary group theory (see, e.g., [7, p. 45]).

THEOREM 4: *Let* $e_1, \ldots, e_n$ *be an* o.n. *basis of* V. *Let* $\pi_1, \ldots, \pi_k$ *be right coset representatives for* G_χ *in* G. *Suppose* $\chi(id)^2 = [G: G_\chi]$. *If* $\alpha \in \Omega$ *is such that* $G_\alpha \subseteq G_\chi$, *then* $\{e^*_{\alpha\pi_i}: 1 \leq i \leq [G: G_\chi]\}$ *is an orthogonal basis of* $\langle e^*_{\alpha\sigma}: \sigma \in G\rangle$.

PROOF. Let $\lambda(\sigma) = \chi(\sigma)/\chi(id)$, $\sigma \in G_\chi$. Since $\alpha \in \Omega$, it follows that λ is identically 1 on G_α.

Now,

$$(e^*_{\alpha\pi_i}, e^*_{\alpha\pi_j}) = \left(e^*_\alpha, e^*_{\alpha\pi_j\pi_i^{-1}}\right)$$

$$= \frac{\chi(id)}{o(G)} \sum_{\tau \in G_\alpha} \chi\left(\pi_i\pi_j^{-1}\tau\right), \text{ from (3)}$$

$$= \frac{\chi(id)}{o(G)} \chi\left(\pi_i\pi_j^{-1}\right) \sum_{\tau \in G_\alpha} \lambda(\tau)$$

$$= \chi(id)\, \chi\left(\pi_i\pi_j^{-1}\right)/[G:\ G_\alpha].$$

The result follows because $\chi(\pi_i\pi_j^{-1}) \neq 0$ if and only if $i = j$ (again appealing to [11]).

3. References

[1] Curtis, Charles, and Reiner, Irving, Representation Theory of Finite Groups and Associative Algebras (Interscience, New York, 1962).

[2] DeMeyer, Frank, Groups with an irreducible character of large degree are solvable, Proc. Amer. Math. Soc. **25**, 615–617 (1970).

[3] DeMeyer, Frank, and Janusz, Gerald, Finite groups with an irreducible representation of large degree, Math. Z. **108**, 145–153 (1969).

[4] Feit, Walter, Characters of Finite Groups (Benjamin, New York, 1967).

[5] Freese, Ralph, Inequalities for generalized matrix functions based on arbitrary characters, Linear Algebra Appl. **7**, 337–345 (1973).

[6] Gagola, S. M., Characters fully ramified over a normal subgroup, preprint.

[7] Herstein, I. N., Topics in Algebra, 2nd. ed., (Xerox, Lexington MA, 1975).

[8] Marcus, Marvin, Finite Dimensional Multilinear Algebra, Part I, (Marcel Dekker, New York, 1973).

[9] Marcus, Marvin, and Minc, Henryk, Generalized matrix functions, Trans. Amer. Math. Soc. **116**, 316–329 (1965).

[10] Merris, Russell, and Pierce, Stephen, Elementary divisors of higher degree associated transformations, Linear and Multilinear Algebra **1**, 241–250 (1973).

[11] Merris, Russell, and Watkins, William, Character induced subgroups, J. Res. Nat. Bur. Stand. (U.S.), **77B** (Math. Sci.), Nos. 3&4, 93–100 (1973).

[12] Pierce, Stephen, Orthogonal decompositions of tensor spaces, J. Res. Nat. Bur. Stand. (U.S.), **74B** (Math. Sci.), No. 1, 41–44 (1970).

[13] Timmer, Kathleen, Subgroups of groups of central type, Trans. Amer. Math. Soc. **189**, 133–162 (1974).

[14] Yellen, Jay, On solvability of certain groups of central type, Notices Amer. Math. Soc. **22**, A–104 (1975).

(Paper 80B2–440)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

The Kronecker Power of a Permutation*

Russell Merris**

(March 22, 1976)

Let G be a permutation group of degree n. Think of the elements, σ, of G as n-square permutation matrices. The paper concerns a reduction of the representation $\sigma \to$ the rth Kronecker power of σ. In case G is the full symmetric permutation group, a formula is given which involves the Stirling numbers of the second kind.

Key words: Bell numbers; branching theorem; Clebsch-Gordon series; irreducible character; matrix functions; multiple transitivity; permutation group; Stirling numbers.

1. Introduction

Let S_n denote the full symmetric permutation group of degree n. For each $\sigma \in G$, let $Q(\sigma)$ be the corresponding permutation matrix, i.e., $Q(\sigma) = (\delta_{i\sigma(j)})$. If G is any subgroup of S_n, then Q is a faithful representation of G whose character, θ, counts the number of fixed points. In this note, we investigate a reduction of θ^r, the character of the rth Kronecker power of Q.

The reduction of the Kronecker (or inner) product of two irreducible representations is called a Clebsch-Gordon series. When $G = S_n$, the problem of obtaining a Clebsch-Gordon series has been solved (see, e.g., [3],[1] [4] or [7]). However, the solution does not easily lead to explicit formulas for the reduction of higher Kronecker powers of representations.

When $1 \leq r \leq n$, the problem naturally arises in connection with a certain class of matrix functions: Let λ be an irreducible character of G. If $A = (a_{ij})$ is an n-square complex matrix, let

$$e_r(A) = \sum_{\sigma \in G} \lambda(\sigma) E_r(a_{1\sigma(1)}, \ldots, a_{n\sigma(n)}),$$

where E_r is the rth elementary symmetric function. It is easily seen, by making special choices for r, G, and λ, that determinant, permanent and trace are all examples of e_r functions. It is proved in [5, Theorem 6] that e_r is not identically zero, if and only if λ is a component of θ^r. Thus, although our general interest is to obtain a reduction of θ^r, we are specifically interested in the smallest number k such that λ is a component of θ^k.

2. Results

If λ and χ are characters of G, then by $(\lambda, \chi)_G$ we denote the usual "inner product" of characters. Let $\Gamma_{r,n}$ be the set of functions from the first r to the first n positive integers. It will sometimes be convenient to think of a function $\gamma \in \Gamma_{r,n}$ as a sequence, $\gamma = (\gamma(1), \gamma(2), \ldots, \gamma(r))$. The rth Kronecker power of $Q(\sigma)$ is an n^r-square matrix which is indexed by $\Gamma_{r,n}$ (usually ordered lexicographically). For $\alpha, \beta \in \Gamma_{r,n}$, the α, β entry of this big matrix is $\prod_{t=1}^{r} \delta_{\alpha(t), \sigma\beta(t)} = \delta_{\alpha, \sigma\beta}$. In

AMS Subject Classification: Primary, 20C30; Secondary, 15A15, 05A19.
*An invited paper. A portion of this work was done during the period when the author was supported by a research grant from the National Science Foundation.
**Present address: Department of Mathematics, California State University, Hayward, CA 94542.
[1] Figures in brackets indicate the literature references at the end of this paper.

particular,

$$\theta^r(\sigma) = \sum_{\gamma \in \Gamma_{r,n}} \delta_{\gamma,\sigma\gamma}.$$

For $\gamma \in \Gamma_{r,n}$, define $G(\gamma) = \{\sigma \in G : \sigma\gamma = \gamma\}$. It is straight forward to verify that $G(\gamma)$ is the subgroup of G which individually fixes the elements of the range of γ.

LEMMA: *Let λ be an irreducible character of G. Then*

$$(\lambda, \theta^r)_G = \sum_{\gamma \in \Gamma_{r,n}} [G: G(\gamma)]^{-1} (\lambda, 1)_{G(\gamma)}. \tag{1}$$

PROOF:

$$\begin{aligned}(\lambda, \theta^r)_G &= \frac{1}{o(G)} \sum_{\sigma \in G} \lambda(\alpha)\theta^r(\sigma) \\ &= \frac{1}{o(G)} \sum_{\sigma \in G} \sum_{\gamma \in \Gamma_{r,n}} \lambda(\sigma)\delta_{\gamma,\sigma\lambda} \\ &= \frac{1}{o(G)} \sum_{\gamma \in \Gamma_{r,n}} \sum_{\sigma \in G(\gamma)} \lambda(\sigma) \\ &= \sum_{\gamma \in \Gamma_{r,n}} [G: G(\gamma)]^{-1} (\lambda, 1)_{G(\gamma)}.\end{aligned}$$

Let $Q_{r,n}$ be the subset of $\Gamma_{r,n}$ consisting of strictly increasing functions, i.e., $\gamma \in Q_{r,n}$ if and only if $1 \leq \gamma(1) < \gamma(2) < \ldots < \gamma(r) \leq n$. Since $G(\gamma)$ consists of those $\sigma \in G$ which individually fix the integers in the range of γ, it follows that $G(\gamma) = \{id\}$ for all $\gamma \in Q_{n-1,n}$. In particular, $(\lambda, 1)_{G(\gamma)} = \lambda(id)$ for all $\gamma \in Q_{n-1,n}$. It follows from the lemma that $(\lambda, \theta^{n-1})_G > 0$ for every irreducible character λ of G. Since the principal (identically one) representation is a component of Q (i.e., $(1, \theta)_G > 0$) we see that $(\lambda, \theta^r)_G \leq (\lambda, \theta^{r+1})_G$. These remarks are restated as

COROLLARY 1: *Let G be a subgroup of S_n. There exists a positive integer $m \leq n-1$ such that $(\chi, \theta^r)_G > 0$ for every irreducible character χ of G and every $r \geq m$.*

In general, we may expect a randomly chosen irreducible character to appear in θ^r for some $r < m$.

COROLLARY 2: *Let λ be an irreducible character of G. Suppose k is the smallest integer such that $(\lambda, \theta^k)_G > 0$. Then*

$$(\lambda, \theta^k)_G = (k!) \sum_{\gamma \in Q_{k,n}} [G: G]^{-1} (\lambda, 1)_{G(\gamma)}. \tag{2}$$

PROOF: Take $\beta \in \Gamma_{k,n}$. Suppose β contains exactly s distinct integers. Let $\alpha \in Q_{s,n}$ be the sequence which contains the distinct integers appearing in β. Then $G(\alpha) = G(\beta)$. If $s < k$, then by (1) and the definition of k, $(\lambda, 1)_{G(\alpha)} = 0$. Therefore, for $r = k$, the only terms of (1) which survive correspond to sequences of distinct integers. These are precisely $\{\gamma\sigma: \gamma \in Q_{k,n}, \sigma \in S_k\}$.

In what follows, we will frequently take $G = S_n$. In this case, we write $(\lambda, \chi)_n$ rather than the more cumbersome $(\lambda, \chi)_{S_n}$.

COROLLARY 3: *Let* λ *be an irreducible character of* S_n. *For* $1 \le r < n$,

$$(\lambda, \theta^r)_n = \sum_{t=1}^{r} S(r, t)(\lambda, 1)_{n-t}, \tag{3}$$

where the numbers $S(r, t)$ *are Stirling numbers of the second kind.*

PROOF: Let γ be a generic element of $\Gamma_{r,n}$. For each $t = 1,2, \ldots, r$, there are $S(r, t)$ ways of partitioning the r components of γ into t nonempty subsets. For each such partition, there are $n!/(n - t)!$ ways of filling the r positions in γ in such a way that the same integer appears in two components if and only if both components belong to the same set of the partition. In other words, there are $n!S(r, t)/(n - t)!$ sequences in $\Gamma_{r,n}$ which contain precisely t distinct integers. Finally, if $G = S_n$, then $G(\gamma)$ consists of the symmetric group on the $n - t$ integers not appearing in γ. The result follows from eq (1).

The numbers

$$b_r = \sum_{t=1}^{r} S(r, t)$$

are called Bell numbers after E. T. Bell. It follows immediately from (3) that for $1 \le r < n$, $(1, \theta^r)_n = b_r$. In fact, a more general result was proved in [6], namely for $1 \le r \le n$, $(1, \theta^r)_G \ge b_r$ with equality if and only if G is r-fold transitive. Strangely, $(1, \theta^{n+1})_G \ge b_{n+1} - 1$, with equality if and only if $G = S_n$.

COROLLARY 4: *Let* λ *be an irreducible character of* S_n. *Let* k *be minimal so that* $(\lambda, \theta^k)_n > 0$. *Then* $(\lambda, \theta^k)_n = (\lambda, 1)_{n-k}$.

PROOF. This is immediate either from Corollary 2 or Corollary 3.

An important feature of Corollary 3 is that we know how to compute $(\lambda, 1)_{n-t}$. The *branching theorem* [2, p. 126] states the following: Suppose λ arises from the frame $(m_1, m_2, \ldots, m_p)$, $m_1 \ge m_2 \ge \ldots \ge m_p$, and $m_1 + m_2 + \ldots + m_p = n$. Then the restriction of λ to S_{n-1} decomposes as $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_p$, where λ_i is the character of S_{n-1} arising from the frame $(m_1, \ldots, m_{i-1}, m_i - 1, m_{i+1}, \ldots, m_p)$, and λ_i is understood not to appear if $m_{i-1} = m_i$. Since the identically one representation of S_{n-t} corresponds to the frame $p = 1$, $m_1 = n - t$, we obtain the following apparently well-known result.

COROLLARY 5: *Let* λ *be the irreducible character of* S_n *which arises from the frame* $(m_1, m_2, \ldots, m_p)$. *Assume that* λ *is not the principal character. The smallest number* k *such that* $(\lambda, \theta^k)_n > 0$ *is* $k = n - m_1$.

EXAMPLE 1: If $\lambda = \epsilon$, the alternating character of S_n, then λ corresponds to the frame $p = n$, $m_1 = m_2 = \ldots = m_n = 1$. In particular, by Corollary 5, $(\epsilon, \theta^r) = 0$ unless $r \ge n - 1$. (This proves that the bound in Corollary 1 is sharp.) Using Corollary 4, we find that $(\epsilon, \theta^{n-1})_n = (\epsilon, 1)_1 = 1$.

EXAMPLE 2: Let $G = S_5$. Let λ be the character arising from the frame (2, 2, 1). From Corollary 5, the smallest k for which $(\lambda, \theta^k)_5 > 0$ is $k = 5 - 2 = 3$. Using Corollary 4 and the branching theorem, we obtain $(\lambda, \theta^3)_5$: Confusing the frame with the character, the restriction of (2, 2, 1) to S_4 is (2, 1, 1) + (2, 2). The further restriction to S_3 is (1, 1, 1) + (2, 1) + (2, 1). Finally, the restriction of (2, 2, 1) to S_2 is 3(1, 1) + 2(2). By Corollary 4, $(\lambda, \theta^3)_5 = (\lambda, 1)_2$. But we have just discovered that $(\lambda, 1)_2 = 2$.

I am grateful to Kenneth R. Rebman for pointing out to me that the numbers in (3) are the Stirling numbers, and to the anonymous referee of an earlier effort for the brief proof of Corollary 3. My original proof, involving the Frobenius Reciprocity Theorem for induced characters, is substantially more complicated; it does not rely on the lemma.

3. References

[1] Berge, C., Principles of Combinatorics (Academic Press, New York-London, 1971).

[2] Boerner, Hermann, Representations of Groups with Special Consideration for the Needs of Modern Physics (North-Holland/American Elsevier, Amsterdam-London/New York, 1970).

[3] Coleman, A. J., Induced Representations with Applications to S_n and $GL(n)$, Queen's Papers in Pure and Applied Math. (Queen's University, Kingston, Ontario, 1966).

[4] Hamermesh, Morton, Group Theory and its Application to Physical Problems, (Addison-Wesley, Reading-London, 1962).

[5] Merris, Russell, Inequalities for matrix functions, J. Algebra **22**, 451–460 (1972).

[6] Merris, Russell, and Pierce, Stephen, The Bell numbers and r-fold transitivity, J. Combinatorial Theory **12**, 155–157 (1972).

[7] Robinson, G., de B. Representation Theory of the Symmetric Group (Univ. of Toronto Press, 1961).

(Paper 80B2–441)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B2, April–June 1976

On Iohvidov's Proofs of the Fischer-Frobenius Theorem*

R. C. Thompson**

(March 12, 1976)

A short proof is given of theorem of Fischer and Frobenius exhibiting a conjunctive transformation mapping Toeplitz matrices onto Hankel matrices.

Key words: Hankel matrix; Toeplitz matrix.

1. Introduction

A Hankel matrix H_{n-1} is an $n \times n$ matrix with the structure

$$H_{n-1} = [s_{i+j}]_{0 \leq i,j<n}.$$

A Toeplitz matrix T_{n-1} is an $n \times n$ matrix with the structure

$$T_{n-1} = [c_{i-j}]_{0 \leq i,j<n}.$$

It is easy to see that Toeplitz matrices may be converted to Hankel matrices in a uniform way by a matrix multiplication. Indeed, if

$$J = \begin{bmatrix} & & & 1 \\ & \bigcirc & 1 & \\ & 1 & \bigcirc & \\ 1 & & & \end{bmatrix},$$

then JT_{n-1} is a Hankel matrix for all Toeplitz matrices T_{n-1}. This procedure, however, does not carry Hermitian Toeplitz matrices to Hermitian (i.e., real) Hankel matrices. The theorem of Fischer-Froebenius asserts that a class of transformations exist each of which uniformly carries Toeplitz matrices to Hankel matrices in such a way that Hermitian Toeplitz matrices are carried to Hermitian Hankel matrices. Recently I. C. Iohvidov has published three proofs of this result. One of these proofs is a direct but somewhat intricate calculation; it may be found on pages 211–213 of [1][1]. A second proof, to be found on page 217 of [1] and also in [2], makes a preliminary reduction to the case of positive definite Toeplitz matrices, then takes advantage of a decomposition of definite Toeplitz matrices known from the theory of the trigonometric moment problem. The third proof, in [2], avoids the reduction to the positive definite case, and uses instead a more complicated decomposition of Toeplitz matrices due to Iohvidov and Krein [3, p. 338].

The purpose of this paper is to give a short and direct proof of the Fischer-Frobenius theorem. Our proof is based on a simple decomposition of arbitrary Toeplitz matrices, for which the proof is almost a triviality and which was apparently not noticed in [1] and [2]. See equation (3). Iohvidov's techniques then may be applied to (3) to produce the desired result rapidly.

*An invited paper.
**Present address: Institute for Applied Algebra, University of California, Santa Barbara, CA 93106.
[1] Figures in brackets indicate the literature references at the end of this paper.

2. A Decomposition of Toeplitz Matrices

Let $\epsilon_0, \epsilon_1, \ldots, \epsilon_{2n-2}$ be distinct fixed numbers on the unit circle, and set

$$V = [\epsilon_j{}^i]_{-(n-1)\leq i\leq n-1,\ 0\leq j\leq 2n-2}.$$

This square matrix V is nonsingular since after removal of a factor from each column it becomes a Vandermonde. Alternatively, a nontrivial linear relation on the rows of V quickly leads to the impossible existence of a nonzero polynomial of degree at most $2n-2$ with $2n-1$ distinct roots.

Let $T = [c_{i-j}]_{0\leq i,j\leq n-1}$ be an arbitrary Toeplitz matrix and take

$$c = [c_{-(n-1)}, \ldots, c_0, \ldots, c_{n-1}]^T$$

to be a column vector formed from the entries of T. Let

$$d = [d_0, \ldots, d_{2n-2}]^T$$

be an unknown column vector. Since V is nonsingular, we may choose d such that

$$c = Vd.$$

That is,

$$c_i = \sum_{t=0}^{n-1} d_t \epsilon_t{}^i + \sum_{t=n}^{2n-2} d_t \epsilon_t{}^i, \qquad -n < i < n. \tag{1}$$

Equations (1) are the same as the matrix equation

$$T_{n-1} = FDF^* + G\Delta G^*, \tag{2}$$

where

$$D = \operatorname{diag}(d_0, \ldots, d_{n-1}),$$

$$\Delta = \operatorname{diag}(d_n, \ldots, d_{2n-2}),$$

$$F = [\epsilon_t{}^i]_{0\leq i\leq n-1,\ 0\leq t\leq n-1},$$

$$G = [\epsilon_t{}^i]_{0\leq i\leq n-1,\ n\leq t\leq 2n-2}.$$

Here F is square, G rectangular, and * denotes conjugate transpose. We shall deduce the Fischer-Frobenius theorem from (2). Note that both FDF^* and $G\Delta G^*$ are Toeplitz matrices.

3. The Fischer-Frobenius Theorem

Take a and b to be fixed complex numbers, with $a\bar{b}$ not real. Let

$$\xi = [\xi_0, \ldots, \xi_{n-1}]^T, \qquad \eta = [\eta_0, \ldots, \eta_{n-1}]^T$$

be column vectors, and λ an indeterminate. If the right-hand side of the polynomial identity

(3) $$\sum_{p=0}^{n-1} \xi_p \lambda^p = \sum_{j=0}^{n-1} (a + \bar{a}\lambda)^{n-1-j} (b + \bar{b}\lambda)^j \eta_j$$

is multiplied out, then compared with the left-hand side, each ξ_p is linearly expressed in the η_j, $j = 0, \ldots, n-1$, with coefficients depending on a and b. Thus we obtain a matrix A for which

(4) $$\xi = A\eta;$$

explicit formulas for the entries of A may be seen in [1, p. 209]. (We shall not need these formulas.) Let $A = [\alpha_{pj}]_{0 \leq p, j \leq n-1}$. This matrix A is nonsingular. Indeed, set $\xi = 0$ in (3) and (4), then replace λ in (3) with $(a\lambda - b)/(\bar{b} - \bar{a}\lambda)$ to obtain

$$\sum_{j=0}^{n-1} \lambda^j \eta_j = 0,$$

after cancelling a factor. Thus $\xi = 0$ implies $\eta = 0$; hence A is nonsingular.

In (3) and (4) set $\eta_0 = \ldots = \eta_{j-1} = \eta_{j+1} = \ldots = \eta_{n-1} = 0$, $\eta_j = 1$, for fixed j. Then, by (4), $\xi_p = \alpha_{pj}$, and hence by (3)

(5) $$\sum_{p=0}^{n-1} \alpha_{pj} \lambda^p = (a + \bar{a}\lambda)^{n-1} \left(\frac{b + \bar{b}\lambda}{a + \bar{a}\lambda} \right)^j.$$

The theorem of Fischer and Frobenius, slightly generalized is this:

THEOREM 1: *Let matrix* A *be defined as above. Then, for any Toeplitz matrix* T_{n-1}, *the matrix*

(6) $$H_{n-1} = A^T T_{n-1} \bar{A}$$

is a Hankel matrix. Conversely, for any Hankel matrix H_{n-1}, *the matrix* T_{n-1} *defined by* (6) *is a Toeplitz matrix.*

PROOF: Choose $\epsilon_0, \ldots, \epsilon_{2n-2}$ on the unit circle, distinct and unequal to $-a/\bar{a}$. We then have the decomposition (2). Since Hankel matrices are closed under addition, it will suffice to prove that both $A^T F D F^* \bar{A}$ and $A^T G \Delta G^* \bar{A}$ are Hankel matrices. We give the proof for the first of these, the proof for the other being similar. Let $f(\lambda) = (b + \bar{b}\lambda)/(a + \bar{a}\lambda)$. This Möbius function f maps the unit circle to the extended real axis, and in particular maps $\epsilon_0, \ldots, \epsilon_{2n-2}$ to finite real values. Set

$$r_i = f(\epsilon_i), \qquad 0 \leqq i \leqq 2n - 2,$$

$$\rho_i = (a + \bar{a}\epsilon_i)^{n-1}, \qquad 0 \leqq i \leqq 2n - 2.$$

Then each r_i is real. The (j, l) element of $A^T F$ is (by (5))

$$\sum_{p=0}^{n-1} \alpha_{pj} \epsilon_l^p = \rho_l f(\epsilon_l)^j = \rho_l r_l^j.$$

Thus

$$A^T F = F_1 D_1$$

where

$$F_1 = [r_l^i]_{0 \le i \le n-1,\ 0 \le l \le n-1}$$

$$D_1 = \text{diag}(\rho_0, \ldots, \rho_{n-1}).$$

Thus F_1 is a real Vandermonde matrix. We now have

$$A^T F D F^* \bar{A} = (A^T F) D (A^T F)^*$$

$$= F_1 (D_1 D D_1^*) F_1^* = F_1 D_2 F_1^T,$$

where $D_2 = D_1 D D_1^*$ is diagonal. However, any matrix of the form $F_1 D_2 F_2^T$ with F_1 a Vandermonde matrix and D_2 diagonal is a Hankel matrix.

Thus the transformation $T_{n-1} \to A^T T_{n-1} \bar{A}$ carries Toeplitz matrices to Hankel matrices. This evidently is a nonsingular linear transformation from the $(2n - 1)$-dimensional complex vector space of Toeplitz matrices into (and therefore onto) the $2n - 1$ dimensional complex space of Hankel matrices. Thus any Hankel matrix H_{n-1} is (uniquely) realizable in the form $A^T T_{n-1} \bar{A}$. This completes the proof.

When T_{n-1} is Hermitian, H_{n-1} is also Hermitian, being a conjuctive transform of T_{n-1}. Conversely, when H_{n-1} is Hermitian, so is its conjunctive transform $T_{n-1} = A^{-1^T} H_{n-1} \bar{A}^{-1}$. Thus we have:

COROLLARY: *Under the Fischer-Frobenius transformation, Hermitian Toeplitz matrices map onto real Hankel matrices, and conversely.*

The preparation of this paper was supported in part by the Air Force Office of Scientific Research, under Grant 72–2164.

3. References

[1] Iohvidov, I. C., Toeplitz and Hankel matrices and forms, Moscow, 1974 (in Russian).

[2] Iohvidov, I. C., On the transformation of Fischer-Frobenius, Theory of functions, functional analysis and its applications, **15** (1972), 203–212 (in Russian).

[3] Iohvidov, I. C., and Krein, M. G., Spectral theory of operators in spaces with an indefinite metric II, American Math. Society Translations, Series 2, vol. **34**, pp. 283–373.

(Paper 80B2–442)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

Matrix Algebra and Eigenvalues for the Bead/Spring Model of Polymer Solutions

Jeffrey T. Fong*

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

and

Anton Peterlin

Institute for Materials Research, National Bureau of Standards, Washington, D.C. 20234

(February 12, 1976)

Based on purely matrix-algebraic arguments, we prove three new results on the eigenvalue problem [Zimm, 1956] arising from modeling the bulk hydrodynamic and dielectric properties of very dilute polymer solutions:

1. Let N be the number of identical segments joining $N+1$ beads of a bead/spring model for any polymer molecule. Then for any N, the well-known hydrodynamic interaction matrix $\underline{\underline{H}}$ is positive definite if the so-called interaction parameter h^* is less than $(2\sqrt{2}+1)/7$ or 0.547.
2. Let $\underline{\underline{A}}$ be the tridiagonal matrix governing the elastic link force on each bead. Then if $\underline{\underline{H}}$ is positive definite, there exists a matrix $\underline{\underline{Q}}$ such that $\underline{\underline{Q}}^{-1}\,\underline{\underline{H}}\,\underline{\underline{A}}\,\underline{\underline{Q}} = \underline{\underline{Q}}^{T}\,\underline{\underline{A}}\,\underline{\underline{Q}} = \underline{\underline{\Lambda}}$, where $\underline{\underline{\Lambda}}$ is a diagonal matrix.
3. (Equivalence Statement due to Lodge and Wu [1971]). The nonzero eigenvalues of the matrix product $\underline{\underline{H}}\,\underline{\underline{A}}$ are completely determined by solving the eigenvalue problem associated with a symmetric matrix $\underline{\underline{S}}$, where $S_{ij} = H_{ij} + H_{i-1,j-1} - H_{i-1,j} - H_{i,j-1}$, $i,j = 1,2,\ldots,N$.

To illustrate the significance of these results, numerical output for N as large as 300 based on a Fortran program for several values of h^* is given.

Key words: Bead/spring model; dilute polymer solutions; eigenvalue; hydrodynamic interaction; matrix algebra; necklace model; polymer physics.

1. Statement of Problem

Some twenty years ago, Zimm [1][1] formulated a linear, second-order partial differential equation for a distribution function ψ which depends on time and $3(N+1)$ coordinates $x_0, y_0, z_0, \ldots, x_N, y_N, z_N$, of the $N+1$ beads, for modeling the bulk behavior of very dilute polymer solutions under the influence of external force, Brownian motion, and hydrodynamic interaction among the beads of the necklace model. The mechanical model for each polymer molecule is that of a chain of N identical, ideally elastic segments joining $N+1$ identical beads with complete flexibility at each bead. Two length parameters are of interest in this model: a_h, the so-called hydrodynamic radius of the bead, and b_0, the root mean square of the segment length. The ratio a of the two length parameters ($a = a_h/b_0$), and the number N of elastic segments completely characterize the mathematical problem

*On leave during 1975–76 as a ComSci Fellow, Nuclear Regulatory Comm., Wash., D.C. 20555.

[1] Figure in brackets indicate the literature references on page 281.

through the following two symmetric, square matrices of order $N+1$:

$$\underline{\underline{A}} = \begin{bmatrix} 1 & -1 & 0 \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 \ldots & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 \ldots & 0 & -1 & 1 \end{bmatrix} \tag{1.1}$$

$$\underline{\underline{H}} = \begin{bmatrix} 1 & a & a.2^{-1} & a.3^{-1} & \ldots\ldots & a.N^{-1} \\ a & 1 & a & a.2^{-1} & \ldots\ldots & a(N-1)^{-1} \\ a.2^{-1} & a & 1 & a & \ldots\ldots & a(N-2)^{-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a.N^{-1} & a(N-1)^{-1} & \ldots\ldots & \ldots\ldots & \ldots\ldots & 1 \end{bmatrix} \tag{1.2}$$

The symmetry of the two matrices, $\underline{\underline{A}}$ and $\underline{\underline{H}}$, and an assumption based on physical arguments that the eigenvalues of the matrix product $\underline{\underline{H}}\,\underline{\underline{A}}$ are distinct, led to an eigenvalue problem with a coordinate transformation matrix $\underline{\underline{Q}}$ such that a similarity transformation with $\underline{\underline{Q}}$ diagonalizes $\underline{\underline{H}}\,\underline{\underline{A}}$, and congruent transformations with $\underline{\underline{Q}}$ diagonalize $\underline{\underline{A}}$ and $\underline{\underline{H}}$ as follows:

$$\underline{\underline{Q}}^{-1}\ \underline{\underline{H}}\,\underline{\underline{A}}\ \underline{\underline{Q}} = \underline{\underline{\Lambda}}\ , \tag{1.3}$$

$$\underline{\underline{Q}}^{T}\ \underline{\underline{A}}\ \underline{\underline{Q}} = \underline{\underline{M}}\ , \tag{1.4}$$

$$\underline{\underline{Q}}^{-1}\,\underline{\underline{H}}\ [\underline{\underline{Q}}^{-1}]^{T} = \underline{\underline{N}} = \underline{\underline{\Lambda}}\,\underline{\underline{M}}^{-1}\ , \tag{1.5}$$

where $\underline{\underline{\Lambda}} = \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_N)$, $\underline{\underline{M}} = \mathrm{diag}(\mu_0, \mu_1, \ldots, \mu_N)$, and $\underline{\underline{N}} = \mathrm{diag}(\nu_0, \nu_1, \ldots, \nu_N)$, are diagonal matrices, and the superscripts $^{-1}$ and t in eqs (1.3) through (1.5) denote the inverse and the transpose of a matrix, respectively. A closed form solution for a specific linear laminar flow with a transverse gradient constant over all beads, was first given by Peterlin and appeared at the end of Zimm's paper [1, p. 278] as a note added in proof. Computationally speaking, the modeling problem is reduced to the construction of the transformation matrix $\underline{\underline{Q}}$ such that both $\underline{\underline{\Lambda}}$ and $\underline{\underline{M}}$ are known for constructing the distribution function ψ.

The purpose of this paper is to derive some new results based on matrix algebraic arguments and another obvious property of the matrix $\underline{\underline{H}}$, namely, its positive-definiteness under certain conditions, which Zimm [1] failed to exploit. In section 2, we first show that $\underline{\underline{H}}$ is positive definite if the parameter a is positive and is bounded from above by a constant $a_{\max}$ which equals 1.0 for $N = 1$, 0.872 for $N = 2$, 0.815 for $N = 3$, and 0.773 for $N \geq 4$. In section 3, we prove our main result

that for positive definite $\underline{\underline{H}}$, there exists some $\underline{\underline{Q}}$ such that $\underline{\underline{\Lambda}} = \underline{\underline{M}}$ and $\underline{\underline{N}} = \underline{\underline{1}}$, i.e., $\lambda_j = \mu_j$ and $\nu_j = 1$ for $j = 0, 1, 2, \ldots, N$. In section 4, we furnish an algebraic proof as an alternate to the proof of a result due to Lodge and Wu [2], which we subsequently use in calculating the eigenvalues of the matrix product $\underline{\underline{H}}\,\underline{\underline{A}}$. In section 5, we discuss the significance of our three mathematical results. To illustrate the significance, we include in an appendix a sample Fortran program and output for several values of the nondimensional parameter a and for N as large as 300.

2. Positive Definiteness of the Interaction Matrix

As shown in (1.2), the so-called interaction matrix $\underline{\underline{H}}$ is not only symmetric, but also a "stripe matrix," i.e., all entries along a line parallel to the main diagonal are equal to one another. Furthermore, $\underline{\underline{H}}$ depends on a scalar parameter a, which is required to be positive ($a > 0$) from physical considerations.[2] We now wish to show that, for any given order of the matrix $\underline{\underline{H}}$, there is an upper bound $a_{\max}$ for the parameter a such that $\underline{\underline{H}}$ is positive definite for $0 < a < a_{\max}$.

Since $\underline{\underline{H}}$ is a stripe matrix, an application of a standard result in matrix algebra[3] implies the corollary that $\underline{\underline{H}}$ is positive definite, if and only if, the determinant of $\underline{\underline{H}}$, i.e., $\det \underline{\underline{H}}$, is positive for a specific order $N + 1$ as well as for all orders less than $N + 1$. It is convenient to introduce a new matrix $\underline{\underline{B}}$ and a new positive constant c as defined below:

$$c = a^{-1} \tag{2.1}$$

$$\underline{\underline{B}} = \begin{bmatrix} c & 1 & 2^{-\frac{1}{2}} & 3^{-\frac{1}{2}} & \cdots\cdots & N^{-\frac{1}{2}} \\ 1 & c & 1 & 2^{-\frac{1}{2}} & \cdot\cdot & (N-1)^{-\frac{1}{2}} \\ 2^{-\frac{1}{2}} & 1 & c & 1 & \cdots\cdots & (N-2)^{-\frac{1}{2}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ N^{-\frac{1}{2}} & (N-1)^{-\frac{1}{2}} & (N-2)^{-\frac{1}{2}} & (N-3)^{-\frac{1}{2}} & \cdots\cdots & c \end{bmatrix} \tag{2.2}$$

Since $\det \underline{\underline{B}} = c^{N+1} \det \underline{\underline{H}}$, $c>0$, our problem is reduced to the study of the positivity of $\det \underline{\underline{B}}$.

From the property of a determinant that its value is unaltered if any multiple of one line is added to a different, parallel line,[4] let us perform the following operations on the matrix $\underline{\underline{B}}$:

Step 1 Keep the first row of $\underline{\underline{B}}$ fixed, and replace the jth row with the difference between the jth and the $(j-1)$th rows for $j = 2,3, \ldots, N+1$.

Step 2 Keep the first column of the new matrix following step 1 fixed, and replace the jth column of the new matrix with the difference between the jth and the $(j-1)$th columns of the new matrix for $j = 2,3, \ldots, N+1$.

---

[2] We have already seen that by definition, $a(=a_s/b_s)$ is required to be positive if the hydrodynamic radius a_s is positive. In Zimm's formulation [1, pp. 270–271], the parameter is defined by the expression $(6\pi^3)^{-\frac{1}{2}} \rho\, \eta^{-1} b^{-1}$, which is positive for $\rho > 0$, $\eta > 0$, and $b > 0$.

[3] See, e.g., Horn [3, pp. 258–260].

[4] See, e.g., Hohn [3, p. 37].

The resulting matrix, to be denoted by $\underline{\underline{B}}'$, is again symmetric as shown below:

(2.3)

$$\underline{\underline{B}}' = \begin{bmatrix} c & 1-c & 2^{-\frac{1}{2}}-1 & \cdots\cdots \\ 1-c & 2(c-1) & -2^{-\frac{1}{2}}+2-c & \cdots\cdots \\ 2^{-\frac{1}{2}}-1 & -2^{-\frac{1}{2}}+2-c & 2(c-1) & \cdots\cdots \\ 3^{-\frac{1}{2}}-2^{-\frac{1}{2}} & -3^{-\frac{1}{2}}+2(2)^{-\frac{1}{2}}-1 & -2^{-\frac{1}{2}}+2-c & \cdots\cdots \\ & & & \cdots\cdots \\ N^{-\frac{1}{2}}-(N-1)^{-\frac{1}{2}} & -N^{-\frac{1}{2}}+2(N-1)^{-\frac{1}{2}}-(N-2)^{-\frac{1}{2}} & & \cdots\cdots \end{bmatrix}$$

Since det $\underline{\underline{B}}'$ = det $\underline{\underline{B}}$, we have further reduced the study of the positive definiteness of the interaction matrix $\underline{\underline{H}}$ to a study of the matrix $\underline{\underline{B}}'$. As it will become clear very shortly, our derivation so far is intended to pave the way for the application of a sufficient condition for the positive definiteness of $\underline{\underline{H}}$ or $\underline{\underline{B}}'$.

We now observe that for the lowest order of $\underline{\underline{H}}$, $\underline{\underline{B}}$, or $\underline{\underline{B}}'$, i.e., all the 2×2 matrices given by (1.2), (2.2), or (2.3), the positivity of det $\underline{\underline{H}}$ (and thus det $\underline{\underline{B}}$ or det $\underline{\underline{B}}'$) always yields the condition that either $a < 1$ or $c > 1$. It follows that the symmetric matrix $\underline{\underline{B}}'$ has the property that all its diagonal elements are positive. A well-known theorem in matrix algebra states that a symmetric matrix with positive diagonal elements is positive definite if the matrix is "strongly diagonal dominant," i.e., the absolute value of the diagonal element at every row exceeds the sum of the absolute values of all off-diagonal elements in that row:[5]

(2.4) $$|B'_{jj}| > \sum_{\substack{k=0,1,\ldots \\ (k\neq j)}}^{N} |B'_{jk}|, \qquad \text{for } j = 0,1,2, \ldots, N.$$

We denote the generic of $\underline{\underline{B}}'$ as B_{jk}, $j,k = 0,1,2, \ldots, N$.

To apply (2.4), we consider two ranges of the values of c for which one can write down easily the $N+1$ inequalities:

Case 1 ($c \geq 2 - 2^{-\frac{1}{2}}$):

(2.5) $j=0$: $\quad c > c - N^{-\frac{1}{2}}$.

(2.6) $j=1$: $\quad 2(c-1) > 2(c-1) - (N-1)^{-\frac{1}{2}} + N^{-\frac{1}{2}}$.

(2.7) $j=2$: $\quad 2(c-1) > 2(c-1) - (N-2)^{-\frac{1}{2}} + (N-1)^{-\frac{1}{2}}$.

. . . . . .

(2.8) $j=N$: $\quad 2(c-1) > c - 1$.

Case 2 ($1 < c < 2 - 2^{-\frac{1}{2}}$):

(2.5) $j=0$: $\quad c > c - N^{-\frac{1}{2}}$.

[5] See, e.g., Schwarz [4, pp. 17–19].

(2.9) $j=1$: $2(c-1) > 1 - 2^{-\frac{1}{2}}$ $(N = 2)$,

(2.10) or, $2(c-1) > 2 - 3(2)^{-\frac{1}{2}} + 3^{-\frac{1}{2}}$ $(N = 3)$,

(2.11) or, $2(c-1) > 2 - 2(2)^{-\frac{1}{2}} - (N-1)^{-\frac{1}{2}} + N^{\frac{1}{2}}$ $(N \geqq 4)$.

(2.12) $j=2$: $2(c-1) > 3 - 2(2)^{-\frac{1}{2}} - c$ $(N = 2)$,

(2.13) or, $2(c-1) > 5 - 3(2)^{-\frac{1}{2}} - 2c$ $(N = 3)$,

(2.14) or, $2(c-1) > 6 - 5(2)^{-\frac{1}{2}} - 2c + 3^{-\frac{1}{2}}$ $(N = 4)$,

(2.15) or, $2(c-1) > 6 - 4(2)^{-\frac{1}{2}} - 2c - (N-2)^{-\frac{1}{2}} + (N-1)^{-\frac{1}{2}}$ $(N \geqq 5)$.

(2.12) $j=3$: $2(c-1) > 3 - 2(2)^{-\frac{1}{2}} - c$ $(N = 3)$,

(2.13) or, $2(c-1) > 5 - 3(2)^{-\frac{1}{2}} - 2c$ $(N = 4)$,

(2.14) or, $2(c-1) > 6 - 5(2)^{-\frac{1}{2}} - 2c + 3^{-\frac{1}{2}}$ $(N = 5$,

(2.15) or, $2(c-1) > 6 - 4(2)^{-\frac{1}{2}} - 2c - (N-3)^{-\frac{1}{2}} + (N-2)^{-\frac{1}{2}}$ $(N \geqq 6)$.

. . . . . .

(2.16) $j=N$: $2(c-1) > 3 - 2(2)^{-\frac{1}{2}} - c$ $(N \geqq 2)$.

What distinguishes the above two cases is the fact that for case 1, every off-diagonal element of $\underline{\underline{B}}'$ is negative, whereas for case 2, care must be exercised on elements B'_{23}, B'_{32}, B'_{34} . . ., $B'_{N,N-1}$, each of which equals $-2^{-\frac{1}{2}} + 2 - c$. An examination of inequalities (2.5) through (2.8) for case 1 implies that they are always satisfied for every N. On the other hand, for case 2, the controlling inequalities are given by (2.9), (2.10), and (2.11). Combining all results derived in this section, including (2.1) where $a = c^{-1}$, we obtain various upper bounds $a_{\max}$ for the parameter a for various N as listed in table 1.

TABLE 1. *Upper bounds for* a *or* h* *for positive definite* $\underline{\underline{H}}$

Number of Chains	$a_{\max}$		$h^*_{\max} = 2^{-1/2}(a_{\max})$
	(Exact)	(Approx.)	(Approx.)
$N = 1$	1	1.0	0.707
$N = 2$	$\frac{1}{17}(12 + 2\sqrt{2})$	0.872	0.617
$N = 3$	$\frac{3216 + 1278\sqrt{2} - 484\sqrt{3} - 288\sqrt{6}}{4273}$	0.815	0.576
$N = 4$ and above	$\frac{1}{7}(4 + \sqrt{2})$	0.773	0.547

We have included in table 1 a list of upper bounds for a parameter h^* which appears frequently in the literature with $h^* = 2^{-\frac{1}{2}} a$. Since most of the applications of the bead/spring model require more than four chains ($N \geq 4$), the key result, $0 < h^* < 0.547$, a criterion which assures positive definiteness of $\underline{\underline{H}}$, will be used in the next section to achieve a significant simplification of Zimm's eigenvalue problem.

3. Diagonalization—A Simplification of Zimm's Result

The derivation of an upper bound for the parameter a ($0 < a < 0.773$) or, equivalently, the parameter h^* ($0 < h^* < 0.547$), with $h^* = 2^{-\frac{1}{2}} a$, as shown in the last section, enables us to inject the positive definiteness of the matrix $\underline{\underline{H}}$ as an additional matrix property into Zimm's eigenvalue problem. We begin with another standard result in matrix algebra that any symmetric, positive definite matrix can be represented by the square of another symmetric matrix.[6] Hence, for $0 < h^* < 0.547$, the symmetric matrix $\underline{\underline{H}}$ is positive definite and there exists a symmetric $\underline{\underline{C}}$ such that $\underline{\underline{H}} = \underline{\underline{C}}^2$, $\underline{\underline{C}}^T = \underline{\underline{C}}$. Since the matrix $\underline{\underline{A}}$ as defined in (1.1) is symmetric, the matrix product $\underline{\underline{C}}\ \underline{\underline{A}}\ \underline{\underline{C}}$ is also symmetric. By the well-known diagonalization theorem on symmetric matrices,[7] we assert that there exist an orthogonal $\underline{\underline{U}}$ and a diagonal $\underline{\underline{D}}$ such that

$$\underline{\underline{C}}\ \underline{\underline{A}}\ \underline{\underline{C}} = \underline{\underline{U}}^{-1}\ \underline{\underline{D}}\ \underline{\underline{U}}; \qquad \underline{\underline{U}}^T\ \underline{\underline{U}} = \underline{\underline{1}}. \tag{3.1}$$

By choosing $\underline{\underline{Q}}$ as the matrix product $\underline{\underline{C}}\ \underline{\underline{U}}^{-1}$, eqs (1.3) through (1.5) can be simplified as follows:

$$\underline{\underline{Q}}^{-1}\ \underline{\underline{H}}\ \underline{\underline{A}}\ \underline{\underline{Q}} = U\ \underline{\underline{C}}^{-1}\ \underline{\underline{C}}^2\ \underline{\underline{A}}\ \underline{\underline{C}}\underline{\underline{U}}^{-1} = \underline{\underline{D}} = \underline{\underline{\Lambda}}, \tag{3.2}$$

$$\underline{\underline{Q}}^T\ \underline{\underline{A}}\ \underline{\underline{Q}} = \underline{\underline{U}}\ \underline{\underline{C}}\ \underline{\underline{A}}\ \underline{\underline{C}}\ \underline{\underline{U}}^{-1} = \underline{\underline{D}} = \underline{\underline{M}}, \tag{3.3}$$

$$\underline{\underline{Q}}^{-1}\ \underline{\underline{H}}\ (\underline{\underline{Q}}^{-1})^T = \underline{\underline{U}}\ \underline{\underline{C}}^{-1}\ \underline{\underline{C}}^2\ \underline{\underline{C}}^{-1}\underline{\underline{U}}^{-1} = \underline{\underline{1}} = \underline{\underline{N}}. \tag{3.4}$$

Thus we have shown that for $0 < h^* < 0.547$, there exists a transformation matrix $\underline{\underline{Q}}$ such that $\underline{\underline{\Lambda}} = \underline{\underline{M}}$, and $\underline{\underline{N}} = \underline{\underline{1}}$.

4. Equivalence of Two Eigenvalue Problems—An Algebraic Proof

For the bead and spring model of dilute polymer solutions, the result of last section shows that it is *sufficient* to find the $N+1$ eigenvalues of the matrix product $\underline{\underline{H}}\ \underline{\underline{A}}$ if the criterion $0 < a < 0.773$ or $0 < h^* < 0.547$ is satisfied. Since $\underline{\underline{A}}$ is singular and all eigenvalues are distinct, det ($\underline{\underline{H}}\ \underline{\underline{A}}$) vanishes and exactly one of those $N + 1$ eigenvalues must be zero, i.e., $\lambda_0 = 0$.

To find the other N eigenvalues, $\lambda_1, \ldots, \lambda_N$, Lodge and Wu [2] introduced an equivalence statement which asserts that the eigenvalue problem for the unsymmetric product $\underline{\underline{H}}\ \underline{\underline{A}}$ of order $N + 1$, aside from the addition of a zero eigenvalue, is equivalent to that of a symmetric matrix

[6] See, e.g., Halmos [2, p. 139].
[7] See, e.g., Halmos [2, p. 157, bottom lines].

$\underline{\underline{S}}$ of order N, where the elements S_{ij} of $\underline{\underline{S}}$ are combinations of elements of $\underline{\underline{H}}$ as follows:

$$S_{ij} = H_{ij} + H_{i-1,j-1} - H_{i-1,j} - H_{i,j-1}, \qquad i,j=1,2,\ldots,N. \tag{4.1}$$

The equivalence statement is important because the matrix $\underline{\underline{S}}$ for the reduced problem is symmetric and, therefore, computationally more attractive than the matrix product $\underline{\underline{H}}\,\underline{\underline{A}}$. To prove that statement, Lodge and Wu [2] introduced a nonsingular matrix $\underline{\underline{A}}_\epsilon$ with elements $(A_\epsilon)_{ij} = A_{ij}$ for each $i,j = 0,1,\ldots,N$, except for $i = j = 0$ where $(A_\epsilon)_{00} = A_{00} + \epsilon^2$. Using a theorem[8] on the continuity of eigenvalues, a triangular decomposition of the matrix $\underline{\underline{A}}_\epsilon$, and a property of the determinant of a matrix product, it was shown that as $\epsilon \to 0$, the nonzero eigenvalues of $\underline{\underline{H}}\,\underline{\underline{A}}_\epsilon$ approach those of $\underline{\underline{H}}\,\underline{\underline{A}}$ and $\underline{\underline{S}}$.

The purpose of this section is to show that Lodge and Wu's equivalence statement is true by a purely algebraic argument without resorting to the continuity theorem involving the parameter ϵ. We first observe that there exists a triangular decomposition of $\underline{\underline{A}}$ such that $\underline{\underline{A}} = \underline{\underline{L}}^T\,\underline{\underline{L}}$:

$$\underline{\underline{L}} = \begin{bmatrix} 0 & 0 & 0 & \cdots & & \\ -1 & 1 & 0 & & & \\ 0 & -1 & 1 & & & \\ & & & & & \\ & & & & 1 & 0 \\ 0 & & & & -1 & 1 \end{bmatrix}. \tag{4.2}$$

Let λ_i be any nonzero eigenvalue of $\underline{\underline{H}}\,\underline{\underline{A}}$, $i = 1,2,\ldots,N$. Let x_i be the corresponding eigenvector of λ_i. By definition, we have

$$\underline{\underline{H}}\,\underline{\underline{A}}\,x_i = \lambda_i x_i. \tag{4.3}$$

Substituting $\underline{\underline{A}} = \underline{\underline{L}}^T\,\underline{\underline{L}}$ into (4.3), we obtain

$$\underline{\underline{H}}\,\underline{\underline{L}}^T\,\underline{\underline{L}}\,x_i = \lambda_i x_i. \tag{4.4}$$

By definition, x_i is a nonzero vector satisfying (4.3), hence the right-hand side of (4.4) cannot vanish. This implies that $\underline{\underline{L}}\,x_i$ is a nonzero vector (proof by contradiction). Therefore it makes sense to premultiply both sides of (4.4) by the singular matrix $\underline{\underline{L}}$:

$$\underline{\underline{L}}\,\underline{\underline{H}}\,\underline{\underline{L}}^T\,\underline{\underline{L}}\,x_i = \lambda_i\,\underline{\underline{L}}\,x_i. \tag{4.5}$$

Denoting $\underline{\underline{L}}\,x_i$ by a nonzero vector y_i, we obtain

$$\underline{\underline{L}}\,\underline{\underline{H}}\,\underline{\underline{L}}^T y_i = \lambda_i y_i. \tag{4.6}$$

[8] Lodge and Wu [2] attribute this theorem to Ostrowski [6].

Equations (4.3) and (4.6) show that the nonzero eigenvalues of $\underline{\underline{H}}\,\underline{\underline{A}}$ and $\underline{\underline{L}}\,\underline{\underline{H}}\,\underline{\underline{L}}^T$ are identical. Some algebraic manipulation with $\underline{\underline{L}}\,\underline{\underline{H}}\,\underline{\underline{L}}^T$, which is symmetric, shows that its nonzero eigenvalues are given by those of $\underline{\underline{S}}$ with components S_{ij} defined in (4.1).

5. Significance of Results

To discuss the significance of our mathematical results, it is appropriate here to introduce a closed-form solution of the distribution function ψ, originally due to Peterlin [1, p. 278 footnote], for a specific linear laminar flow with a constant transverse velocity gradient κ expressible through a nondimensional gradient parameter β:[9]

$$\psi = \psi_1 \psi_2 \ldots\ldots \psi_i \ldots\ldots \psi_N, \tag{6.1}$$

where

$$\psi_i = C_i \exp\left\{-\frac{3}{2}\frac{\mu_i}{1+\beta^2/\lambda_i^2}\left[\xi_i^2 - \frac{2\beta\xi_i\zeta_i}{\lambda_i} + \left(1+\frac{2\beta^2}{\lambda_i^2}\right)\zeta_i^2 + \left(1+\frac{\beta^2}{\lambda_i^2}\right)\eta_i^2\right]\right\}, \tag{6.2}$$

$C_i = (\mu_i/\pi)^{3/2}(1+\beta^2/\lambda_i^2)^{-1/2}$, and λ_i, μ_i, $i = 1,2, \ldots, N$, are the nonzero diagonal elements of matrices $\underline{\underline{\Lambda}}$ and $\underline{\underline{M}}$ as defined in (1.3) and (1.4), respectively. The symbols ξ_i, ζ_i, η_i in (6.2) denote the nondimensionalized transformed coordinates of the $(i+1)$th bead as defined below:

$$\xi_i = \sum_{j=1}^{N} b_0^{-1}\,[Q^{-1}]_{ij}\,x_j. \tag{6.3}$$

$$\zeta_i = \sum_{j=1}^{N} b_0^{-1}\,[Q^{-1}]_{ij}\,y_j. \tag{6.4}$$

$$\eta_i = \sum_{j=1}^{N} b_0^{-1}\,[Q^{-1}]_{ij}\,z_j. \tag{6.5}$$

The principal result of this paper has been the demonstration that the diagonal elements of both matrices $\underline{\underline{\Lambda}}$ and $\underline{\underline{M}}$ are identical for each i, $i = 0,1,2, \ldots, N$, if the hydrodynamic interaction parameter h^* is less than 0.547. This means that once we have computed the values of λ_i, $i = 0,1,2, \ldots, N$, from the eigenvalue problem of the matrix product $\underline{\underline{H}}\,\underline{\underline{A}}$ as defined in (1.3), it is no longer necessary, as understood previously by Zimm [1], to compute Q explicitly in order to calculate the diagonal elements μ_i, $i = 0,1,2, \ldots, N$, of the matrix $\underline{\underline{M}}$ as defined in (1.4).

The result $\lambda_i = \mu_i$, $i = 0,1,2, \ldots, N$, for $h^* < 0.547$ is also significant in discrediting some computational results in the literature such as Zimm, Roe, and Epstein [7], etc., where the transformation matrix $\underline{\underline{Q}}$, being not explicitly known, was expeditiously approximated by the orthogonal transformation matrix $\underline{\underline{Q}}_R$ of the well-known Rouse model of zero hydrodynamic interaction (i.e., $\underline{\underline{H}} = \underline{\underline{1}}$). Under that circumstance, the matrix $\underline{\underline{M}}$ is given by $\underline{\underline{Q}}_R^T\underline{\underline{A}}\,\underline{\underline{Q}}_R$ or

[9] The flow field is specified by the velocity vector $\mathbf{v} = (\kappa y, 0, 0)$, and the parameter β is related to κ by the formula $\beta = \pi\eta_0 b_0^2\,\kappa/(kT)$, where k is Boltzman's constant and T is the absolute temperature.

[10] See, e.g., Zimm, Roe and Epstein [7], Hearst [8], and Tschoegl [9].

$\underline{\underline{Q}}_k^{-1} \underline{\underline{A}} \underline{\underline{Q}}_k$, and the eigenvalues μ_j are given by $4 \sin^2 [j \pi/2(N+1)]$ which contradict our exact result with $\lambda_j = \mu_j$, and $\underline{\underline{M}}^{-1} \underline{\underline{A}} = \underline{\underline{N}} = \underline{\underline{1}}$ as shown in (3.4).

Computationally speaking, we are interested in the solution of the eigenvalue problem (4.3), for large N, and particularly, in the smaller values of λ_i. The equivalence statement of Lodge and Wu [2] provides the basis for computing those eigenvalues very economically with the aid of a large-scale computer. The exact computer-generated calculation renders many earlier approximation schemes obsolete.[10] Without the benefit of the equivalence statement, Thurston and Morrison [10] used a computer to obtain the eigenvalues of the matrix product $\underline{\underline{H}} \underline{\underline{A}}$ for values of N up to 15. Based on their equivalence statement, Lodge and Wu [11] were able to extend the computational capability and reported some numerical results for bead/spring models with as many as 300 links. Unfortunately, Lodge and Wu did not exhibit all eigenvalues, nor the computer program with which eigenvalues could be generated for interesting applications. In the appendix we list a Fortran program and output to illustrate the significance of this paper and to enable others to extend these calculations, if they so desire.

We are indebted to Morris Newman, Senior Research Mathematician, National Bureau of Standards, Washington, D.C., for substantial original contribution and critical comments during the course of this investigation. We also wish to thank Martin Cordes of NBS Applied Mathematics Division for technical consultation related to the computational aspect of this work.

6. References

[1] Zimm, B. H., J Chem. Phys. **24**, 269 (1956).
[2] Lodge, A. S., and Wu, Y. -j., Rheol. Acta, **10**, 539 (1971)
[3] Hohn, F. E., Elementary Matrix Algebra (MacMillan, 1958).
[4] Schwarz, H. R., Numerical Analysis of Symmetric Matrices, translated from the German by P. Hertelendy (Prentice-Hall, 1973).
[5] Halmos, P. R., Finite-Dimensional Vector Spaces (Van Nostrand, 1958).
[6] Ostrowski, A., Jber. Deutsch. Math. Verein. **60**, 40 (1957).
[7] Zimm, B. H., Roe, G. M., and Epstein, L. F., J. Chem. Phys. **24**, 279, (1956).
[8] Hearst, J. E., J. Chem. Phys. **37**, 2547 (1962).
[9] Tschoegl, N. W., J. Chem. Phys. **39**, 149 (1963).
[10] Thurston, G. B., and Morrison, J. D., Polymer **10**, 421 (1969).
[11] Lodge, A. S., and Wu, Y. -j., University of Wisconsin—Mathematics Research Center Report No. 1250. Madison, Wisconsin (1972).

7. Appendix

The following is a Fortran V driver program as executed on a UNIVAC–1108 Exec 8 computer at National Bureau of Standards Laboratory at Gaithersburg, Md. To simplify our programming effort, we take advantage of the well-documented subroutines furnished by the International Mathematical and Statistical Libraries, Inc., of Houston, Texas. For a complete description of these subroutines used in our program, the reader may consult the reference manual entitled "IMSL Library 2, Edition 4 (FORTRAN V), UNIVAC 1100 Series, Reference Manual, November 1974." The appearance of trade names such as UNIVAC and IMSL does not imply an endorsement of their product or services.

```
1*          COMMON A(300),B(300),D(300),E(300),E2(300),HMINOR(45150)
2*              N = 300
3*              HSTAR = 0.4
4*          DO 20 I = 1,N
5*              A(I) = HSTAR * SQRT (2./FLOAT(I))
6*    20    CONTINUE
7*              B(I) = 1. - A(I)
8*          DO 30 I = 2,N
9*              B(I) = A(I - 1) - A(I)
10*   30    CONTINUE
11*         DO 50 I = 1,N
12*             II = I * (I+1) / 2
13*             HMINOR(II) = 2. * B(I)
14*   50    CONTINUE
15*         N1 = N - 1
16*         DO 60 I = 1, N1
17*             N2 = N - I
18*             DO 70 J = 1, N2
19*                 K = ((I+J) * (I+J-1) / 2) + J
20*                 HMINOR(K) = B(I+1) - B(I)
21*   70        CONTINUE
22*   60    CONTINUE
23*         CALL EHOUSS (HMINOR, N, D, E, E2)
24*         CALL EQRT2S (D, E, N, DUMMY, 1, IERR)
25*         IF (IERR .NE. 0) GO TO 100
26*         SUMD1 = 0
27*         SUMD2 = 0
28*         DO 80 I = 1,N
29*             SUMD1 = SUMD1 + 1./D(I)
30*             SUMD2 = SUMD2 + 1./D(I)/D(I)
31*   80    CONTINUE
32*  200    FORMAT (1H1, 10X, 4HN = , I4, 20X, 8HHSTAR =      ,F6.2,//)
33*         WRITE (6,200) N, HSTAR
34*  250    FORMAT (1H, 4X, 42HNONZERO EIGENVALUES FOR MATRIX PRODUCT H*A,//)
35*         WRITE (6,250)
36*  320    FORMAT (1H0, 10F12.5)
37*         WRITE (6,320) (D(I), I=1, N)
38*  330    FORMAT (1H0,8X,E12.5,2X,20H(SUM OF RECIPROCALS))
39*         WRITE (6,330) SUMD1
40*  340    FORMAT (1H0,8X,E12.5,2X,27H(SUM OF RECIPROCAL SQUARES))
41*         WRITE (6,340) SUMD2
42*  100    WRITE (6,400) IERR
43*  400    FORMAT (1H1, 10X, 36HROUTINE FAILS AT EIGENVALUE NUMBER  ,I4)
44*         END
```

For $N = 300$, and $h^* = 0.4$ (corresponding to $a = 0.566$), the following output was obtained at a total run cost of \$47.91 for a total run time of 5 min 42 s. The CPU cost was \$24.21 with a run time of 5 min 36 s.

```
N = 300        HSTAR = .40
NONZERO EIGENVALUES FOR MATRIX PRODUCT H*A
 .11539-02  .35945-02  .67180-02  .10383-01  .14480-01  .18953-01  .23749-01  .28837-01  .34185-01  .39772-01
 .45574-01  .51576-01  .57762-01  .64119-01  .70634-01  .77297-01  .84098-01  .91028-01  .98078-01  .10524+00
 .11251+00  .11988+00  .12734+00  .13489+00  .14252+00  .15023+00  .15801+00  .16586+00  .17377+00  .18174+00
 .18976+00  .19784+00  .20596+00  .21412+00  .22233+00  .23058+00  .23886+00  .24717+00  .25551+00  .26387+00
 .27226+00  .28067+00  .28909+00  .29754+00  .30599+00  .31446+00  .32294+00  .33142+00  .33991+00  .34840+00
 .35690+00  .36539+00  .37388+00  .38236+00  .39084+00  .39932+00  .40778+00  .41623+00  .42467+00  .43309+00
 .44150+00  .44990+00  .45827+00  .46663+00  .47496+00  .48328+00  .49157+00  .49983+00  .50807+00  .51629+00
 .52447+00  .53263+00  .54075+00  .54885+00  .55691+00  .56495+00  .57294+00  .58091+00  .58883+00  .59673+00
 .60458+00  .61240+00  .62017+00  .62791+00  .63531+00  .64326+00  .65088+00  .65845+00  .66598+00  .67347+00
 .68091+00  .68830+00  .69566+00  .70296+00  .71022+00  .71743+00  .72459+00  .73171+00  .73878+00  .74580+00
 .75277+00  .75969+00  .76656+00  .77338+00  .78015+00  .78686+00  .79353+00  .80014+00  .80670+00  .81321+00
 .81967+00  .82608+00  .83243+00  .83872+00  .84497+00  .85116+00  .85730+00  .86338+00  .86941+00  .87538+00
 .88130+00  .88716+00  .89297+00  .89873+00  .90443+00  .91008+00  .91567+00  .92120+00  .92668+00  .93211+00
 .93748+00  .94279+00  .94806+00  .95326+00  .95841+00  .96351+00  .96855+00  .97353+00  .97847+00  .98334+00
 .98816+00  .99293+00  .99765+00  .10023+01  .10069+01  .10115+01  .10160+01  .10204+01  .10248+01  .10291+01
 .10334+01  .10376+01  .10418+01  .10460+01  .10500+01  .10541+01  .10580+01  .10619+01  .10658+01  .10696+01
 .10734+01  .10771+01  .10808+01  .10844+01  .10880+01  .10915+01  .10949+01  .10984+01  .11017+01  .11051+01
 .11083+01  .11116+01  .11147+01  .11179+01  .11210+01  .11240+01  .11270+01  .11299+01  .11328+01  .11357+01
 .11385+01  .11413+01  .11440+01  .11467+01  .11493+01  .11519+01  .11545+01  .11570+01  .11595+01  .11619+01
 .11643+01  .11666+01  .11689+01  .11712+01  .11734+01  .11756+01  .11778+01  .11799+01  .11820+01  .11840+01
 .11860+01  .11880+01  .11899+01  .11918+01  .11937+01  .11955+01  .11973+01  .11991+01  .12008+01  .12025+01
 .12041+01  .12058+01  .12073+01  .12089+01  .12104+01  .12119+01  .12134+01  .12148+01  .12163+01  .12176+01
 .12190+01  .12203+01  .12216+01  .12229+01  .12241+01  .12253+01  .12265+01  .12277+01  .12288+01  .12299+01
 .12310+01  .12321+01  .12331+01  .12341+01  .12351+01  .12361+01  .12370+01  .12379+01  .12388+01  .12397+01
 .12435+01  .12414+01  .12422+01  .12430+01  .12438+01  .12445+01  .12452+01  .12460+01  .12466+01  .12473+01
 .12480+01  .12486+01  .12492+01  .12492+01  .12504+01  .12510+01  .12515+01  .12521+01  .12526+01  .12531+01
 .12536+01  .12541+01  .12545+01  .12550+01  .12554+01  .12558+01  .12562+01  .12566+01  .12570+01  .12573+01
 .12577+01  .12580+01  .12583+01  .12586+01  .12589+01  .12592+01  .12595+01  .12597+01  .12600+01  .12602+01
 .12600+01  .12606+01  .12608+01  .12610+01  .12612+01  .12614+01  .12615+01  .12617+01  .12618+01  .12619+01
 .12670+01  .12621+01  .12622+01  .12623+01  .12674+01  .12624+01  .12625+01  .12625+01  .12625+01  .12625+01
```

By changing the first three statements of the original program, one can find the eigenvalues for any N and h^* as long as the core storage is not exceeded and the value of h^* is less than 0.547. To illustrate this point, we show below for $N = 250$ and $h^* = 0.3$ the first three statements of the modified program and the appropriate output. The total cost for the run was \$25.29 with a total run time of 3 min 19 s. The CPU cost was \$13.96 and the CPU time was 3 min 14 s.

```
1*          COMMON A(250),B(250),D(250),E(250),E2(250),HMINOR(31375)
2*            N = 250
3*          HSTAR = 0.3
```

N = 250 HSTAR = .30

NONZERO EIGENVALUES FOR MATRIX PRODUCT H*A

.11737−02	.36804−02	.69245−02	.10767−01	.15099−01	.19865−01	.25015−01	.30516−01	.36336−01	.42455−01
.48849−01	.55504−01	.62402−01	.69531−01	.76877−01	.84432−01	.92184−01	.10012+00	.10824+00	.11653+00
.12499+00	.13360+00	.14237+00	.15128+00	.16033+00	.16951+00	.17882+00	.18826+00	.19781+00	.20748+00
.21726+00	.22714+00	.23713+00	.24721+00	.25739+00	.26766+00	.27802+00	.28846+00	.29898+00	.30958+00
.32025+00	.33099+00	.34180+00	.35268+00	.36362+00	.37462+00	.38568+00	.39679+00	.40795+00	.41916+00
.43042+00	.44173+00	.45308+00	.46446+00	.47589+00	.48735+00	.49894+00	.51037+00	.52192+00	.53350+00
.54511+00	.55674+00	.56039+00	.58006+00	.59174+00	.60345+00	.61516+00	.62689+00	.63863+00	.65038+00
.66213+00	.67389+00	.68565+00	.69742+00	.70918+00	.72094+00	.73270+00	.74446+00	.75621+00	.76795+00
.77969+00	.79141+00	.80312+00	.81482+00	.82650+00	.83817+00	.84982+00	.86145+00	.87306+00	.88465+00
.89622+00	.90776+00	.91928+00	.93077+00	.94224+00	.95368+00	.96508+00	.97646+00	.98780+00	.99912+00
.10104+01	.10216+01	.10328+01	.10440+01	.10551+01	.10662+01	.10773+01	.10883+01	.10993+01	.11102+01
.11211+01	.11319+01	.11427+01	.11534+01	.11641+01	.11748+01	.11854+01	.11959+01	.12064+01	.12168+01
.12272+01	.12376+01	.12478+01	.12581+01	.12682+01	.12783+01	.12884+01	.12984+01	.13083+01	.13182+01
.13280+01	.13378+01	.13475+01	.13571+01	.13667+01	.13762+01	.13856+01	.13950+01	.14043+01	.14136+01
.14228+01	.14319+01	.14409+01	.14499+01	.14588+01	.14676+01	.14764+01	.14851+01	.14938+01	.15023+01
.15108+01	.15192+01	.15276+01	.15358+01	.15440+01	.15521+01	.15602+01	.15682+01	.15761+01	.15839+01
.15916+01	.15993+01	.16069+01	.16144+01	.16218+01	.16292+01	.16365+01	.16437+01	.16508+01	.16579+01
.16648+01	.16717+01	.16785+01	.16853+01	.16919+01	.16905+01	.17050+01	.17114+01	.17177+01	.17239+01
.17301+01	.17362+01	.17421+01	.17481+01	.17539+01	.17596+01	.17653+01	.17709+01	.17764+01	.17818+01
.17871+01	.17923+01	.17975+01	.18025+01	.18075+01	.18124+01	.18172+01	.18220+01	.18266+01	.18312+01
.18356+01	.18400+01	.18443+01	.18485+01	.18526+01	.18567+01	.18606+01	.18645+01	.18683+01	.18719+01
.18755+01	.18790+01	.18825+01	.18858+01	.18891+01	.18922+01	.18953+01	.18983+01	.19012+01	.19040+01
.19067+01	.19093+01	.19119+01	.19143+01	.19167+01	.19190+01	.19212+01	.19233+01	.19253+01	.19272+01
.19290+01	.19308+01	.19324+01	.19340+01	.19355+01	.19368+01	.19381+01	.19393+01	.19405+01	.19415+01
.19424+01	.19433+01	.19440+01	.19447+01	.19453+01	.19458+01	.19462+01	.19465+01	.19467+01	.19469+01

(Paper 80A2–443)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

Predictable Regular Continued Cotangent Expansions*

Jeffrey Shallit**

(March 22, 1976)

Expansions of the form $x = \cot(\text{arc cot } n_0 - \text{arc cot } n_1 + \text{arc cot } n_2 - \ldots)$ are discussed. It is shown that if x is of the form $\frac{1}{2}(c + \sqrt{c^2 + 4})$, then the n's are predictable by a simple recurrence.

Continued fractions derived from the expansion of x are also given.

Key words: Continued cotangent; continued fraction; quadratic irrational.

1. Introduction

In "A Cotangent Analogue of Continued Fractions" [1][1], D. H. Lehmer discussed expansions of the form

$$x = \cot(\text{arc cot } n_0 - \text{arc cot } n_1 + \text{arc cot } n_2 - \ldots). \tag{1}$$

This expansion is called a *continued cotangent*. The expansion is called a *regular continued cotangent* if

(a) n_s is a positive integer satisfying $n_s \geq n_{s-1}^2 + n_{s-1} + 1$ $(s = 1, 2, \ldots)$.

(b) If the expansion (1) is finite and n_k is the last n, then $n_k > n + n_{k-1}^2 + n_{k-1} + 1$.

Given any positive real number x, its regular continued cotangent expansion is generated by the following algorithm:

$$x_0 = x, \qquad n_0 = [x_0] \tag{2}$$

$$x_{s+1} = \frac{x_s n_s + 1}{x_s - n_s}, \; n_{s+1} = [x_{s+1}] \qquad (s = 1, 2, \ldots). \tag{3}$$

As usual, the brackets denote the greatest integer function.

Lehmer called the x_s's *complete cotangents* and the n_s's *incomplete* or *partial cotangents*.

He did not find any combination of familiar constants whose regular continued cotangent expansion was in any way predictable.

Here we present an infinite sequence of quadratic irrationals with the property that each member of the sequence has a predictable regular continued cotangent expansion.

AMS Subject Classification: Primary, 10A30; Secondary, 10F20, 10A25.

*An invited paper.

**Present address: Princeton University, Princeton, N.J., 08540.

[1] Figures in brackets indicate the literature references at the end of this paper.

2. Irrationals With Predictable Expansions

We will show that, for a positive integer c,

$$\frac{c + \sqrt{c^2 + 4}}{2}$$

has a predictable regular continued cotangent expansion. More precisely, we state the following:

THEOREM: *Let α,β be the roots of the quadratic $x^2 - cx - 1 = 0$, with c a positive integer, and $\alpha > \beta$. Let α be expanded into its regular continued cotangent expansion, with $n_0, n_1, n_2, \ldots$ being the sequence of partial cotangent, and $x_0, x_1, x_2, \ldots$ being the sequence of complete cotangents. Then*

$$\text{(I)} \qquad x_k = \alpha^{3^k}$$
$$\text{(II)} \qquad n_k = \alpha^{3^k} + \beta^{3^k}.$$

PROOF: We start out by stating some facts about α and β:

$$\alpha = \frac{c + \sqrt{c^2 + 4}}{2}, \quad \beta = \frac{c - \sqrt{c^2 + 4}}{2}. \tag{4}$$

This follows from the definition of α and β.

$$\alpha\beta = -1, \qquad \alpha + \beta = c, \qquad \alpha - \beta = \sqrt{c^2 + 4}, \ [\alpha] = c. \tag{5}$$

We now define the sequence V_k:

$$V_k = \alpha^k + \beta^k \qquad (k \geq 0) \tag{6}$$

It is easily shown by induction that

$$V_k = cV_{k-1} + V_{k-2} \qquad (k \geq 2). \tag{7}$$

We are now ready to prove the theorem. The proof of part (I) proceeds by induction.

A. Verification for $k = 0$:

$$x_0 = \alpha^{3^0} = \alpha^1 = \alpha$$
$$n_0 = \alpha^{3^0} + \beta^{3^0} = \alpha^1 + \beta^1 = c$$
$$= [x_0] = [\alpha] = c.$$

B. Assume the theorem is true for $k = s$. Then we want to show that the theorem is true for $k = s + 1$.

From (3), we have

$$x_{s+1} = \frac{x_s n_s + 1}{x_s - n_s}$$

$$= \frac{(\alpha^{3^s})(\alpha^{3^s} + \beta^{3^s}) + 1}{\alpha^{3^s} - (\alpha^{3^s} + \beta^{3^s})}$$

$$= \frac{\alpha^{2\cdot 3^s} + \alpha^{3^s}\beta^{3^s} + 1}{-\beta^{3^s}}$$

$$= \frac{\alpha^{2\cdot 3^s}}{-\beta^{3^s}} \qquad (\text{since } \alpha\beta = -1)$$

$$x_{s+1} = \alpha^{3^{s+1}} \qquad (\text{since } \beta = -1/\alpha)$$

The proof of part (I) of the theorem is now complete by induction. We complete the proof of the theorem by showing that

$$n_{s+1} = [x_{s+1}] = [\alpha^{3^{s+1}}] = \alpha^{3^{s+1}} + \beta^{3^{s+1}}$$

Since $\beta = -1/\alpha$, and $\alpha > 1$, we have

$$-1 < \beta^r < 0 \qquad (r \text{ odd}, \geq 1). \tag{8}$$

If we add α^r to this inequality, we get

$$\alpha^r - 1 < \alpha^r + \beta^r < \alpha^r \qquad (r \text{ odd}, \geq 1) \tag{9}$$

Now it is obvious that α^r is a quadratic irrational, so neither $\alpha^r - 1$ nor α^r are integers. Therefore, there must be exactly one integer between $\alpha^r - 1$ and α^r. Since $V_0 = \alpha^0 + \beta^0 = 2$ and $V_1 = \alpha^1 + \beta^1 = c$, we have, from (7), that $V_r = \alpha^r + \beta^r$ is an integer for $r \geq 0$. Therefore, V_r *is the integer* between $\alpha^r - 1$ and α^r, and we have

$$[\alpha^r] = \alpha^r + \beta^r \qquad (r \text{ odd}, \geq 1). \tag{10}$$

We may now take $r = 3^{s+1}$ to get

$$n_{s+1} = [x_{s+1}] = [\alpha^{3^{s+1}}] = \alpha^{3^{s+1}} + \beta^{3^{s+1}} \qquad (s \geq 0).$$

The proof of both parts of the theorem is now complete.

In the following tables, we give the values of α and β for the first few values of c, and the values of n_k for the first few values of c and k.

TABLE 1. *Values of* α *and* β

c	α	β
1	$\frac{1}{2}(1 + \sqrt{5})$	$\frac{1}{2}(1 - \sqrt{5})$
2	$1 + \sqrt{2}$	$1 - \sqrt{2}$
3	$\frac{1}{2}(3 + \sqrt{13})$	$\frac{1}{2}(3 - \sqrt{13})$
4	$2 + \sqrt{5}$	$2 - \sqrt{5}$
5	$\frac{1}{2}(5 + \sqrt{29})$	$\frac{1}{2}(5 - \sqrt{29})$

TABLE 2. *Values of* n_k

k \ c	1	2	3	4	5
1	1	2	3	4	5
2	4	14	36	76	140
3	76	2786	46764	439204	2744420

3. Some Observations on n_k and x_k

First, we note that, as a special case of the theorem for $c = 1$, we have $\alpha = \phi$, the golden ratio, and $V_k = L_k$, the kth Lucas number [2]. In fact, we have

$$x_k = \phi^{3^k} \text{ and } n_k = L_{3^k}.$$

Second, from part (II) of the theorem, it is not difficult to show that

$$n_{k+1} = n_k^3 + 3n_k \qquad (k \geq 0). \tag{11}$$

We also point out that, empirically, the regular continued cotangent expansion of an "average" irrational number satisfies

$$n_{k+1} = O(n_k^2)$$

For $\alpha = \frac{1}{2}(c + \sqrt{c^2 + 4})$, however, we have, in view of (11)

$$n_{k+1} = O(n_k^3)$$

so that, in a certain sense, this group of quadratic irrationals is approximated unusually well by the continued cotangent algorithm.

Third, we observe that the regular continued cotangent for $1/\alpha = -\beta = \frac{1}{2}(-c + \sqrt{c^2 + 4})$ also is predictable. For if the regular continued cotangent expansion of x $(x > 1)$ is $n_0, n_1, n_2, \ldots$ then the expansion for $1/x$ is $0, n_0, n_1, n_2, \ldots$ From this it easily follows that the expansion for $1/\alpha$ is predictable as follows:

$$x_0 = 1/\alpha, \qquad n_0 = 0$$

$$x_k = \alpha^{3^{k-1}}, \qquad n_k = \alpha^{3^{k-1}} + \beta^{3^{k-1}} \qquad (k \geq 1).$$

We now introduce the sequence U_k, defined as follows:

$$U_k = \frac{\alpha^k - \beta^k}{\alpha - \beta}. \tag{12}$$

The sequence U_k satisfies the same recurrence as V_k, that is,

$$U_k = cU_{k-1} + U_{k-2} \qquad (k \geq 2). \tag{13}$$

Now $x_k = \alpha^{3^k}$ is a quadratic irrational and can obviously be put in the form

$$x_k = v_k + u_k\sqrt{c^2 + 4} \tag{14}$$

where v_k and u_k are rational numbers. From the definitions of V_k and U_k, it is easily verified that

$$v_k = \frac{1}{2}V_{3^k} \tag{15}$$

$$u_k = \frac{1}{2}U_{3^k}. \tag{16}$$

Letting $m_k = U_{3^k}$, we have the following recurrence formula which can be verified by substitution:

$$m_{k+1} = (c^2 + 4)m_k^3 - 3m_k \tag{17}$$

$$v_{k+1} = 4v_k^3 + 3v_k. \tag{18}$$

$$u_{k+1} = 4(c^2 + 4)u_k^3 - 3u_k. \tag{19}$$

Lehmer [3] observed that

$$2u_k = (n_0^2 + 1)(n_1^2 + 1)(n_2^2 + 1)\ldots(n_{k-1}^2 + 1) \tag{20}$$

We also have

$$(v_{k+1})/v_k = (u_{k+1})/u_k + 2. \tag{21}$$

Many similar identities can be obtained.

4. Unusual Continued Fractions

We observe that the *regular* continued fractions for α and $1/\alpha$ are as follows:

$$\alpha = c + \frac{1}{c} + \frac{1}{c} + \frac{1}{c} + \frac{1}{c} + \ldots$$

$$1/\alpha = \frac{1}{c} + \frac{1}{c} + \frac{1}{c} + \frac{1}{c} + \ldots.$$

Lehmer showed that if the n_k's (the partial cotangents) for a real number x are known, then x can be expanded into the following irregular continued fraction:

$$x = n_0 + \frac{n_0^2 + 1}{n_1 - n_0} + \frac{n_1^2 + 1}{n_2 - n_1} + \frac{n_2^2 + 1}{n_3 - n_2} + \ldots.$$

It can be shown by induction that the kth convergent to this continued fraction, p_k/q_k satisfies

$$p_k/q_k = U_e/U_{e-1} \qquad \text{where } e = (3^k + 1)/2 \tag{22}$$

The first few such continued fractions are as follows:

$$\frac{1}{2}(1+\sqrt{5}) = 1 + \frac{2}{3} + \frac{17}{72} + \frac{5777}{439128} + \cdots$$

$$1+\sqrt{2} = 2 + \frac{5}{12} + \frac{197}{2772} + \frac{7761797}{21624369228} + \cdots$$

$$\frac{1}{2}(3+\sqrt{13}) = 3 + \frac{10}{33} + \frac{1297}{46728} + \frac{2186871697}{102266868085272} + \cdots.$$

These continued fractions converge extremely rapidly.

5. References

[1] Lehmer, D. H., A Cotangent Analogue of Continued Fractions, Duke Math. J. Vol. 4 (1938), pps. 323–340.
[2] Hoggatt, Verner E. Jr., Fibonacci and Lucas Numbers (1969).
[3] Lehmer, D. H., personal communication.

(Paper 80B2–444)

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

Automatic Computing Methods for Special Functions. Part III. The Sine, Cosine, Exponential Integrals, and Related Functions

Irene A. Stegun and Ruth Zucker

Institute for Basic Standards, National Bureau of Standards, Washington, D.C. 20234

(April 29, 1976)

Accurate, efficient, automatic methods for computing the sine, cosine, exponential integrals and hyperbolic sine and cosine integrals are detailed and implemented in an American National Standard FORTRAN program. The functions are also tabulated to 35 significant figures for arguments 0, $10^J\,(10^J)\,10^{J+1}$ with $J = -2(1)2$.

Key words: Continued fraction; cosine integral; exponential integral; FORTRAN program; hyperbolic sine and cosine integrals; key values; recurrence relations.

1. Introduction

Since the sine, cosine, exponential integrals and hyperbolic sine, and cosine integrals are frequently encountered together in physical problems and their expansions have terms in common, we have incorporated these functions into Part III. (For Parts I and II, see.[1]).

While accuracy over the entire domain of definition remains our main concern, we have tended toward methods that also ensure efficiency, portability, and ease of programming and modification. The number of terms in series, the number of convergents in an iterative process, the starting arguments for different methods, are all determined by the program as a function of word length, arguments, accuracy desired, etc. More realistic results are returned when error conditions are encountered. The proper analytic behavior of the function will always be retained to further ensure correct limiting values, in particular of related functions and for purposes of differentiation and integration.

In Parts I and II in addition to the implementing ANS FORTRAN program, we had included a driver (test) program and its results. Since either of these driver programs can be readily modified to compute other functions, we have omitted the driver program and in place of its results have included a table of correct results to 35 significant figures covering essentially the functional range of present computers.

2. Mathematical Properties

Relevant formulas are collected here for completeness and ease of reference. In keeping with the convention of the Handbook [1],[2] x here is a real variable.

[1] Automatic Computing Methods for Special Functions. Part I. Error, Probability, and Related Functions. J. Res. Nat. Bur. Stand. (U.S.), **74B**, (Math. Sci), No. 3, 211–224 (July–Sept. 1970). Automatic Computing Methods for Special Functions. Part II. The Exponential Integral $E_n(x)$. J. Res. Nat. Bur. Stand. (U.S.), **78B**, (Math. Sci), No. 4, 199–216 (Oct.–Dec. 1974).

[2] Figures in brackets indicate the literature references on page

A. Definitions

$$Si(x) = \int_0^x \frac{\sin t}{t}\,dt$$

$$Ci(x) = \gamma + \ln x + \int_0^x \frac{\cos t - 1}{t}\,dt$$

$$Ei(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t}\,dt = \int_{-\infty}^{x} \frac{e^{t}}{t}\,dt \qquad (x > 0)$$

(For $\int_x^\infty \frac{e^{-t}}{t}\,dt = E_1(x)$, often denoted by $-Ei(-x)$, see Part II.)

$$Shi(x) = \int_0^x \frac{\sinh t}{t}\,dt \qquad = \frac{Ei(x) + E_1(x)}{2}$$

$$Chi(x) = \gamma + \ln x + \int_0^x \frac{\cosh t - 1}{t}\,dt = \frac{Ei(x) - E_1(x)}{2}$$

γ (Euler's constant) = 0.57721 56649 . . ..

B. Series Expansions

$$Si(x) = \sum_{k=0}^{\infty} \frac{(-1)^k\, x^{2k+1}}{(2k+1)(2k+1)!}$$

$$Ci(x) = \gamma + \ln x + \sum_{k=1}^{\infty} \frac{(-1)^k\, x^{2k}}{(2k)(2k)!}$$

$$Ei(x) = \gamma + \ln x + \sum_{k=1}^{\infty} \frac{x^k}{(k)(k)!} \qquad (x > 0)$$

$$Shi(x) = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)(2k+1)!}$$

$$Chi(x) = \gamma + \ln x + \sum_{k=1}^{\infty} \frac{x^{2k}}{(2k)(2k)!}$$

C. Continued Fraction

$$-Ci(x) + i[Si(x) - \pi/2] = e^{-ix}\left[\frac{1}{ix+}\,\frac{1}{1+}\,\frac{1}{ix+}\,\frac{2}{1+}\,\frac{2}{ix+\ldots}\right] \qquad (0 < x)$$

$$= E_1(ix)$$

D. Asymptotic Expansions

$$Si(x) = \pi/2 - f(x)\cos x - g(x)\sin x$$

$$Ci(x) = f(x)\sin x - g(x)\cos x$$

where $f(x) \sim \frac{1}{x}\sum_{k=0}^{\infty}\frac{(-1)^k(2k)!}{x^{2k}}$

and $g(x) \sim \frac{1}{x}\sum_{k=0}^{\infty}\frac{(-1)^k(2k+1)!}{x^{2k+1}}$

$$Ei(x) \sim \frac{e^x}{x}\sum_{k=0}^{\infty}\frac{k!}{x^k} \qquad (x > 0)$$

$$Shi(x) = \frac{1}{x}[p(x)\cosh x + q(x)\sinh x]$$

$$Chi(x) = \frac{1}{x}[p(x)\sinh x + q(x)\cosh x]$$

where $p(x) \sim \sum_{k=0}^{\infty}\frac{(2k)!}{x^{2k}}$

and $q(x) \sim \sum_{k=0}^{\infty}\frac{(2k+1)!}{x^{2k+1}}$

E. Special Values

$$Si(0) = 0 \qquad Ci(0) = -\infty \qquad Ei(0) = -\infty$$
$$Shi(0) = 0 \qquad Chi(0) = -\infty$$

F. Symmetry Relations

$(x > 0)$

$$Si(-x) = -Si(x) \qquad Ci(-x) = Ci(x) - i\pi$$
$$Shi(-x) = -Shi(x) \qquad Chi(-x) = Chi(x) - i\pi$$

G. Interrelations

$$Si(x) = \frac{1}{2i}[E_1(ix) - E_1(-ix)] + \pi/2$$

$$Ci(x) = -\frac{1}{2}[E_1(ix) + E_1(-ix)]$$

$$Ei(x) = -\frac{1}{2}[E_1(-x + i0) + E_1(-x - i0)] \qquad (x > 0)$$

H. Value at Infinity

$$\lim_{x \to \infty} Si(x) = \pi/2$$

I. Related Function
Logarithmic Integral

$$li(x) = Ei(\ln(x)) \qquad (x > 1)$$

3. Method

Evaluation of the integrals by means of quadrature formulas suited to the particular type of integrand tends to be inefficient and inaccurate. For $Si(x)$ and $Ci(x)$, the use of the asymptotic expansion is not valid for moderate values of x, while the use of the continued fraction is inefficient and also inaccurate for small values of x. An examination of the series expansion for the functions indicates several difficulties. Summation of the alternating series for $Si(x)$ and $Ci(x)$ will lead to greater round-off errors as x increases. The partial sum at a particular value of k may be zero. Additionally there may be cancellations in adding the logarithmic term and/or Euler's constant for $Ci(x)$, $Ei(x)$ and $Chi(x)$. The more rapidly accumulating round-off errors, in particular when summations are limited to a single register, eliminate the prolonged use of the series expansion. Since the maximum of $Ci(x)$ occurs at $\pi/2$ and $Si(x) = \pi/2$ at $x \simeq 1.92$, testing indicates $x = 2$(PSLSC) as a reasonable starting point for the use of the continued fraction. The starting point for the valid use of the asymptotic expansion for $Si(x)$ and $Ci(x)$ does not coincide with the starting point for $Ei(x)$ ($Shi(x)$ and $Chi(x)$). Testing also indicates that fewer terms are needed in the continued fraction than in the asymptotic expansion.

The following table gives an indication of the number of terms needed to obtain maximum machine accuracy for particular values of x with the various methods of computation. Throughout the paper, NBM is the maximum number of binary digits in the mantissa of a floating point number, and TOLER $= 2^{-NBM}$.

Method	Number of Terms					
	NBM = 27			*NBM* = 60		
$x = 2$	$Si(x)$	$Ci(x)$	$Ei(x)$	$Si(x)$	$Ci(x)$	$Ei(x)$
Power Series	7	7	13	12	12	23
Continued Fraction (Even Form)	24	24	—	106	106	—
Numerical Integration (Trapezoidal or Simpson's Rule)	64	128	64	—	—	—
$x = 24$						
Power Series	—	—	55	—	—	80
Asymptotic Expansion	[a]7,11	7,11	13	—	—	—
Continued Fraction	5	5	—	14	14	—
Numerical Integration	512	512	1024	—	—	—
$x = 48$						
Power Series	—	—	—	—	—	119
Asymptotic Expansion	4,5	4,5	8	16,23	16,23	31
Continued Fraction	4	4	—	9	9	—
$x = 88$						
Asymptotic Expansion	3,4	3,4	6	9,10	9,10	17
Continued Fraction	5	5	—	8	8	—

[a] We indicate the number of odd and even terms of the respective series.

The most accurate, efficient, automatic methods for $Si(x)$ and $Ci(x)$ then are the power series and the continued fraction; for $Ei(x)$, $Shi(x)$ and $Chi(x)$ the power series and the asymptotic

expansion. The lower limit (AELL) for the use of the asymptotic expansion may be shown to approximate |ln TOLER|=NBM(ln 2), where TOLER is the requested upper limit for the relative error. With this choice of the lower limit, one can also show that $Shi(x) \simeq Chi(x) \simeq \frac{1}{2} Ei(x)$. It is necessary then to consider only the asymptotic expansion for $Ei(x)$.

The series computations have been so arranged that the maximum number of functions may be obtained in a minimum of time. The even and odd terms of the series are summed independently both with and without the factor $(-1)^k$. Since $Si(x) - \pi/2$ and $Ci(x)$ are the imaginary and real parts respectively of the continued fraction expansion for $E_1(ix)$, there would be a saving in computing time with options on the functions to be computed. Invalid results are initially supplied for all functions. With the parameter $IC = 1$, $Si(x)$ and $Ci(x)$ only are computed; with $IC = 2$, $Ei(x)$ and $e^{-x}Ei(x)$ only; with $IC = 3$, $Ei(x)$, $e^{-x}Ei(x)$, $Shi(x)$ and $Chi(x)$ only and with $IC = 4$, all functions are computed.

The implementing program checks the input parameters. If IC is outside the range 1–4. the working indicator IND is automatically set equal to 4. Since $Ei(x)$ is defined for positive x only, if $IC = 2$ and $x < 0$, there is an error return and the indicator IERR is set equal to 1. If $x < 0$, $IC = 3$, $Shi(x)$ and $Chi(x)$ are computed; for $IC = 4$, $Si(x)$ and Ci(x) are also computed, invalid results are returned for Ei(x) and $e^{-x}Ei(x)$ and the indicator IERR is set equal to 1. For $x > 0$, the indicator IERR is set equal to zero and valid results are returned only for the functions requested by the parameter IC (or IND).

The computations are performed for positive x (=T) only. Before the return, use is then made of the symmetry relations. Various cases are treated independently. Among these are $x = 0$ and x equal to or greater than the supplied upper limit (ULSC) for the sine and cosine routine. The appropriate values of the functions are supplied. To avoid unnecessary computations of the exponential function and possible overflows and underflows in the final results, if T is equal to or less than the upper limit for the relative error, the exponential of half the argument is set equal to 1. If T is equal to or greater than the computed limiting argument (XMAXHF) for $Shi(x)$ and $Chi(x)$, the maximum machine value (RINF) is supplied for the exponential of half the argument as well as for $Shi(x)$ and $Chi(x)$. The computed limiting argument (XMAXEI) for $Ei(x)$ can be shown to be approximately ln RINF + ln[ln RINF + ln(ln RINF)] $-1/$ln RINF. The value of XMAXHF is approximately XMAXEI + ln 2. The computation of $e^{x/2}$ provides a slight improvement in accuracy and an extension of the range of x. Throughout the program, overflows are avoided as well as underflows affecting accuracy. Other underflows are assumed to be set equal to zero.

For $|x| \leq$ PSLSC (=2), all functions are computed by means of the power series. For $|x| >$ PSLSC, $Si(x)$ and $Ci(x)$ are computed by means of the continued fraction. Only if $IC = 4$ is the working indicator IND set equal to 3. The functions $Ei(x)$, $e^{-x}Ei(x)$, $Shi(x)$ and $Chi(x)$ are then computed by means of the power series or the asymptotic expansion depending on whether $|x| \leq$ AELL or $>$ AELL respectively. With NBM = 27, AELL $\approx$ 18.7 and with NBM = 60, AELL $\approx$ 41.6. To avoid underflow, $|x|$ is tested against a lower limit argument PSLL(=2$\sqrt{}$ AMIN). To simplify computation, AMIN, a minimum machine value is computed as the reciprocal of the maximum machine value (RINF). If $|x| \leq$ PSLL, only the first term of the series of odd terms is used.

The following series definitions are in use

$Si(x)$ = SI = SUMS = $\sum$ SGN (RK)*$TM(RK)$ $\quad IP = -1$ (RK = 1,3, . . .)

$Ci(x)$ = CI = SUMC + XLOG + EULER

where SUMC = $\sum$ SGN (RK)*$TM(RK)$ $\quad IP = 1$ (RK = 2,4, . . .)

$Ei(x)$ = EI = SUMET + SUMOT + XLOG + EULER

$Shi(x)$ = SHI = SUMOT = $\sum TM(RK)$ $\quad IP = -1$ (RK = 1,3, . . .)

$Chi(x)$ = CHI = SUMET[= $\sum TM(RK)$] + XLOG + EULER $\quad IP = 1$ (RK = 2,4,. . .)

with SGN(1) = 1, SGN(RK + 1) = $-$SGN(RK) for RK = 1,3,. . ., and SGN(RK + 1) = SGN(RK) for RK = 2,4, ... The term $TM(RK) = [T^k/k!]/k$ = PTM(RK)/RK where PTM(1) = T and PTM(RK + 1) = PTM(RK)[T/(RK + 1)], $RK \geq 1$.

The series of even and odd terms are always computed together. If the relative error RE computed as $TM/|SUM|$ is less than the prescribed tolerance both series are considered to have converged. If IND = 1 or 4, SUM is replaced by SUMS or SUMC; otherwise by SUMET or SUMOT. To avoid underflow, in generating the terms for $|x| \leq 2$, if PTM $\leq$ AMIN $(RK)^2/T$, the series are likewise considered to have converged. If the sum of terms is zero, the relative RE is automatically set equal to the maximum machine value. This condition is not encountered if the power series for $Si(x)$ and $Ci(x)$ is restricted to the region $|x| \leq 2$. It has been retained to permit the program's use for experimental purposes.

To enable the continued fraction computations to be performed in double precision, since complex quantities are involved, the real notation only is used. Testing has also confirmed the improved accuracy and efficiency of this course. The continued fraction for $Si(x)$ and $Ci(x)$ in its "even" form

$$E_1(ix) = -Ci(x) + i[Si(x) - \pi/2] = e^{-ix}\left[\frac{1}{1+ix} - \frac{1}{3+ix} - \frac{4}{5+ix} - \ldots\right]$$
$$e^{-ix}[F]$$

is evaluated in the forward direction. The first convergent $F_1/G_1 = A_1/B_1$ where $A_1 = 1$, $A_M = -(M-1)^2$, $B_M = 2M - 1 + ix$. If we define

$$F_{-1} = 1,\ F_0 = 0,\ G_{-1} = 0 \text{ and } G_0 = 1$$

then successive convergents F_M/G_M for $M = 1,2, \ldots$ may be obtained by the following recurrence relation

$$F_M = B_M F_{M-1} + A_M F_{M-2}$$
$$G_M = B_M G_{M-1} + A_M G_{M-2}$$

The continued fraction is considered to have converged either if in effect the relative error is equal to or less than the prescribed tolerance or the relative error increases.

Since the successive convergents are complex, $(RE)^2$ is compared with $(TOLER)^2$ where $(RE)^2 = \left[\text{mod}\left(1 - \frac{F_{M-1}/G_{M-1}}{F_M/G_M}\right)\right]^2$. Throughout the computation, to avoid overflow, there is scaling by the absolute maximum (=TMAX) of the real and imaginary parts of the numerator and denominator of the successive convergents. In addition, there is scaling by |TMAX| if the product of the real part of $(B_M - A_M)$ and |TMAX| is equal to or greater than 1/4 the maximum machine value.

The successive terms of the asymptotic expansion are likewise obtained by recurrence with $T_0 = 1$ and $T_K = [K/T]T_{K-1}$ for $K \geq 1$. Since the sum of terms for $Ei(x)$ is always greater than one, the term itself is a good approximation to the relative error. The summation is terminated when a term is less than the prescribed tolerance or the term is equal to or greater than the preceding term. In the latter case, the preceding term is subtracted from the summation to minimize the error.

4. Range

The range for $Si(x)$ and $Ci(x)$ (as well as the accuracy) is limited to the range (and accuracy) of the sine and cosine routine ($|x| <$ ULSC). For the UNIVAC 1108, namely, $x < 2^{21}$ in single precision and $x < 2^{56}$ in double precision. For the function $Ei(x)$, the range of x is essentially the range of the exponential routine. The function $Ei(x)$ is set equal to the machine maximum (RINF) for x beyond XMAXEI, approximately 92.5 in single precision and 715.6 in double precision. For the function $e^{-x}Ei(x)$ beyond $x =$ ULSC only the first two terms of the asymptotic expansion are used. The functions $Shi(x)$ and $Chi(x)$ are set equal to the maximum machine value for x beyond XMAXHF, approximately 93.2 in single precision and 716.3 in double precision.

5. Accuracy and Precision

Using the UNIVAC 1108 to compute the functions, the maximum relative error, except for regions in the immediate neighborhood of zeros, is 4.5 (−7) for single precision computations and 7.5 (−17) for double precision computations. Various auxiliary functions are available to greater accuracy at intermediate points in the subroutine. For example, since $\underset{x\to\infty}{Si}(x) \to \pi/2$, $Si(x) - \pi/2$ should be taken as the imaginary part of the continued fraction. The functions $Ci(x)$, $Ei(x)$ and $Chi(x) - \gamma - \ln x$ are available from the sum of the appropriate series.

The precision may be set lower than the maximum by varying the value of NBM or deleting NBM and setting a precomputed value of TOLER. The above relative errors give an indication of the allowance for round-off errors.

6. Timing—UNIVAC 1108 Time/Sharing Executive System

(The time estimates given below are highly dependent on the operating system environment and consequently should not be relied on for critical timing measurements.)

Single Precision $NBM = 27$		Double Precision $NBM = 60$		
For $Si(x)$ and $Ci(x)$				
Region	Time (seconds)	Region	Time (seconds)	Method
0(0.01)2 (201 values)	0.40	0(.01)2 (201 values)	0.98	Power Series
2(.5)100 (197 values)	.56	2(.5)100 (197 values)	2.06	Continued Fraction
Maximum Time/Evaluation	.0023		0.0059	Power Series
($x = 2$)	.0093	($x = 2$)	.070	Continued Fraction
For $Ei(x)$				
0(.1)18 (181 values)	0.54	0(.2)41 (206 values)	2.05	Power Series
18(.1)41 (231 values)	.28			Asymptotic Expansion
41(.25)100 (237 values)	.25	41(.25)100 (237 values)	0.70	Asymptotic Expansion
Maximum Time/Evaluation	.0044		.016	Power Series
($x = 18$)	.0015	($x = 41$)	.0040	Asymptotic Expansion

7. Testing

The double precision results obtained were compared against available published values. Check values were obtained, where appropriate, by overlapping the power series with either the asymptotic expansion or the continued fraction. Various forms of the continued fraction were also employed as well as numerical integration. Two multi-precision packages[3] were also utilized with varied precision. The single and double precision results agreed with the multi-precision results within the reported accuracy.

8. Many-Place Tables

In the appendix, we have included three tables; one for $Si(x)$ and $Ci(x)$, one for $Ei(x)$ and $e^{-x}Ei(x)$ and one for $Shi(x)$ and $Chi(x)$. The functions are tabulated to 35 significant figures for $x = 0, 10^J (10^J) 10^{J+1}$ with $J = -2(1)2$.

[3] (Private Communication) Peavy, Bradley A. A Multi-Precision Arithmetic Package for Use with the UNIVAC 1108.
Wyatt, W. T. Jr., Lozier, D. W. and Orser, D. J., A Portable Extended-Precision Arithmetic Package and Library with FORTRAN Precompiler, ACM TOMS, Sept. 1975.

9. Special Values

Zeros

$Si(x_s) = \pi/2$

$x_0 = 1.92644\ 7660$

$x_1 = 4.89383\ 5953$

$x_2 = 7.97268\ 2624$

$Ci(x_s) = 0$

$x_0 = 0.61650\ 5486$

$x_1 = 3.38418\ 0423$

$x_2 = 6.42704\ 7744$

$Ei(x) = 0$

$x = 0.37250\ 74107\ 81366\ 63446\ 19918\ 66580\ 11913$

$Chi(x) = 0$

$x = 0.52382\ 25713\ 89864$

Maxima

$Si(\pi) = 1.85193\ 70519\ 82466$

$Ci(\pi/2) = 0.47200\ 06514\ 39569$

Minima

$Si(2\pi) = 1.41815\ 15761\ 32628$

$Ci(3\pi/2) = -0.19840\ 75606\ 92358$

Related Constants

$$\sum_{N=0}^{\infty} \frac{(-1)^N}{(2N+1)(2N+1)!} = Si(1) = 0.94608\ 30703\ 67183\ 01494\ 13533\ 13823\ 17965$$

$$\sum_{N=1}^{\infty} \frac{(-1)^N}{(2N)(2N)!} = Ci(1) - \gamma = -0.23981\ 17420\ 00564\ 72594\ 38658\ 86193\ 25166$$

$$Ci(1) = .33740\ 39229\ 00968\ 13466\ 26462\ 03889\ 15076$$

$$\sum_{N=1}^{\infty} \frac{1}{(N)(N)!} = Ei(1) - \gamma = 1.31790\ 21514\ 54403\ 89486\ 00088\ 44249\ 23183$$

$$Ei(1) = 1.89511\ 78163\ 55936\ 75546\ 65209\ 34331\ 63426$$

$$\sum_{N=0}^{\infty} \frac{1}{(2N+1)(2N+1)!} = Shi(1) = 1.05725\ 08753\ 75728\ 51457\ 18423\ 54895\ 87795$$

$$\sum_{N=1}^{\infty} \frac{1}{(2N)(2N)!} = Chi(1) - \gamma = 0.26065\ 12760\ 78675\ 38028\ 81664\ 89353\ 35387$$

$$Chi(1) = .83786\ 69409\ 80208\ 24089\ 46785\ 79435\ 75630$$

γ(Euler's constant) $= 0.57721\ 56649\ 01532\ 86060\ 65120\ 90082\ 40243$

$\pi/2 = 1.57079\ 63267\ 94896\ 61923\ 13216\ 91639\ 75144$

$\pi = 3.14159\ 26535\ 89793\ 23846\ 26433\ 83279\ 50288$

$\log_e 2 = 0.69314\ 71805\ 59945\ 30941\ 72321\ 21458\ 17656$

2^{-24} = 0.59604 64477 53906 25 (−7)
2^{-27} = .74505 80596 92382 8125 (−8)
2^{-36} = .14551 91522 83668 51806 64062 5 (−10)
2^{-48} = 35527 13678 80050 09293 55621 33789 0625 (−14)
2^{-56} = .13877 78780 78144 56755 29539 58511 35253 90625 (−16)
2^{-60} = .86736 17379 88403 54720 59622 40695 95336 91406 25 (−18)
2^{-108} = .30814 87911 01957 73648 89564 70813 58837 09660 96263
71446 21112 38390 20729 06494 14062 5 (−32)

$\log_e(2^{-24})$ = −16.63553 23334 38687 42601 35709 14996 23763
$\log_e(2^{-27})$ = −18.71497 38751 18523 35426 52672 79370 76733
$\log_e(2^{-36})$ = −24.95329 85001 58031 13902 03563 72494 35645
$\log_e(2^{-48})$ = −33.27106 46668 77374 85202 71418 29992 47526
$\log_e(2^{-56})$ = −38.81624 21113 56937 32736 49988 01657 88781
$\log_e(2^{-60})$ = −41.58883 08335 96718 56503 39272 87490 59408
$\log_e(2^{-108})$ = −74.85989 55004 74093 41706 10691 17483 06935

Maximum and Minimum Machine Values and Their Natural Logarithms
NBC = Number of binary digits in the (biased) characteristic of a floating point number

$$2^{-(2^{NBC-1}+1)} \leq x < 2^{2^{NBC-1}-1}$$

NBC = 8

2^{127} = 0.17014 11834 60469 23173 16873 03715 88410 (39)
2^{-129} = .14693 67938 52785 93849 60920 67152 78070 (−38)
$\log_e(2^{127})$ = 88.02969 19311 13054 29598 84794 25188 42414
$\log_e(2^{-129})$ = −89.41598 62922 32944 91482 29436 68104 77728

NBC = 11

2^{1023} = 0.89884 65674 31157 95386 46525 95394 51236 (308)
2^{-1025} = .27813 42323 13400 17288 62790 89666 55050 (−308)
$\log_e(2^{1023})$ = 709.08956 57128 24051 53382 84602 51714 62914
$\log_e(2^{-1025})$ = −710.47586 00739 43942 15266 29244 94630 98227

10. References

[1] Abramowitz, M. and Stegun, I. A., Handbook of Mathematical Functions, Nat. Bur. Stand. (U. S.), Appl. Math. Ser. 55, (1964).
[2] Blanch, G., Numerical evaluation of continued fractions, SIAM Review **6,** 4, 383–421(1964).
[3] British Association for the Advancement of Science, Mathematical Tables, Vol. 1, 3d ed. (Cambridge University Press, Cambridge, England, 1951).
[4] Glaisher, J. W. L., Tables of the numerical values of the sine-integral, cosine-integral and exponential-integral, Phil. Trans. Roy. Soc. London **160,** 367–388(1870).
[5] Harris, F. E., Tables of the exponential integral $Ei(x)$, MTAC **11,** 9–16(1957).
[6] Lee, Kin L., High-precision Chebyshev series approximation to the exponential integral, NASA TN D–5953, 1970.
[7] Miller, James and Hurst, R. P., Simplified calculation of the exponential integral, MTAC **12,** 187–193(1958).
[8] Murnaghan, Francis D., and Wrench, John W. Jr., The Converging Factor for the Exponential Integral, Report 1535, David Taylor Model Basin, 1963.
[9] National Bureau of Standards, Table of sine and cosine integrals for arguments from 10 to 100, Appl. Math. Ser. **32,** (1954).
[10] National Bureau of Standards, Tables of Sine, Cosine and Exponential integrals, Vol. I & II, MT5 and MT6, (1940).

```
1*      C                              APPENDIX
2*      C
3*      C                          IMPLEMENTING PROGRAM
4*      C LANGUAGE. AMERICAN NATIONAL STANDARD FORTRAN
5*      C DEFINITIONS. X, A REAL VARIABLE
6*      C     SI(X) =INTEGRAL(SIN T/T)DT FROM 0 TO X
7*      C     SI(-X)=-SI(X)
8*      C     CI(X) =GAMMA+LN X+INTEGRAL((COS T-1)/T)DT FROM 0 TO X
9*      C     CI(-X)=CI(X)-I PI
10*     C     EI(X) =-P.V.INTEGRAL(EXP(-T)/T)DT FROM -X TO INFINITY
11*     C     EXNEI(X)=EXP(-X)*EI(X)                          (X .GT. 0)
12*     C         INTEGRAL(EXP(-T)/T) DT FROM X TO INFINITY, OFTEN
13*     C         DENOTED BY -EI(-X)=E1(X). (SEE AUTOMATIC COMPUTING
14*     C         METHODS FOR SPECIAL FUNCTIONS, PART II. THE EXPO-
15*     C         NENTIAL INTEGRAL EN(X), J. OF RESEARCH NBS, 78B,
16*     C         OCTOBER-DECEMBER 1974, PP. 199-216.)
17*     C     SHI(X) =INTEGRAL(SINH T/T)DT FROM 0 TO X
18*     C     SHI(-X)=-SHI(X)
19*     C     CHI(X)=GAMMA+LN X+INTEGRAL((COSH T-1)/T)DT FROM 0 TO X
20*     C     CHI(-X)=CHI(X)-I PI
21*     C                  GAMMA(EULER'S CONSTANT)=.5772156649...
22*     C   SPECIAL CASES
23*     C     X=0
24*     C       SI(0)=SHI(0)=0
25*     C       CI(0)=EI(0)=EXNEI(0)=CHI(0)=-INFINITY
26*     C                                  =-MAX. MACH. VALUE (RINF)
27*     C     LIMITING VALUES - X APPROACHES INFINITY
28*     C       SI(X)=PI/2
29*     C       CI(X)=0
30*     C       EI(X)=SHI(X)=CHI(X)=INFINITY (RINF)
31*     C       EXNEI(X)=0
32*     C USAGE. CALL SICIEI (IC,X,SI,CI,CII,EI,EXNEI,SHI,CHI,CHII,
33*     C                                                      IERR)
34*     C     FORMAL PARAMETERS
35*     C         IC       INTEGER TYPE                        INPUT
36*     C                      IC  FUNCTIONS TO BE COMPUTED
37*     C                       1    SI,CI
38*     C                       2    EI,EXNEI
39*     C                       3    EI,EXNEI,SHI,CHI
40*     C                       4    SI,CI,EI,EXNEI,SHI,CHI
41*     C         X        REAL OR DOUBLE PRECISION TYPE       INPUT
42*     C         SI=SI(X)              (SAME TYPE AS X)      OUTPUT
43*     C         CI+I CII=CI(X)              ''              OUTPUT
44*     C         EI=EI(X)                    ''              OUTPUT
45*     C         EXNEI=EXP(-X)*EI(X)         ''              OUTPUT
46*     C         SHI=SHI(X)                  ''              OUTPUT
47*     C         CHI+I CHII=CHI(X)           ''              OUTPUT
48*     C         IERR     INTEGER TYPE                       OUTPUT
49*     C                      IERR=0   X .GE. 0, NORMAL RETURN
50*     C                      IERR=1   X .LT. 0, ERROR RETURN IF
51*     C                                          IC=2
52*     C MODIFICATIONS.
53*     C     THE CODE IS SET UP FOR DOUBLE PRECISION COMPUTATION
54*     C     WITH DOUBLE PRECISION TYPE STATEMENTS
55*     C          DOUBLE PRECISION FUNCTION REFERENCES AND,PARTICU-
56*     C     LARLY,FOR THE UNIVAC 1108 WITH (SEE DEFINITIONS BELOW)
57*     C          RINF APPROX. 2**1023,ULSC=2**56,NBM=60 AND OTHER
58*     C     CONSTANTS IN DOUBLE PRECISION FORMAT TO 19 SIGNIFICANT
```

```
C     FIGURES. ALL ABOVE ITEMS MUST BE CHANGED FOR SINGLE
C     PRECISION COMPUTATIONS WITH DATA ADJUSTMENTS FOR OTHER
C     COMPUTERS.
C   AUXILIARY FUNCTIONS
C       VARIOUS FUNCTIONS ARE AVAILABLE TO GREATER ACCURACY
C       AT INTERMEDIATE POINTS IN THE SUBROUTINE,NAMELY,
C           SI-(PI/2)=IMAG. PART OF THE CONTINUED FRACTION
C           CI(EI AND CHI)-GAMMA-LN X=SUM OF SERIES
C   CAUTION - THE SUBROUTINE CANNOT READILY BE ADAPTED TO
C             COMPUTE THE FUNCTIONS FOR COMPLEX ARGUMENTS.
C METHOD.   T=ABS(X)
C     POWER SERIES   T .LE. PSLSC(=2) FOR SI,CI
C                    T .LE. AELL(=-LN(TOLER)) FOR EI,SHI,CHI
C         SI=SUMS(SGN(RK)*TM(RK))  IP=-1  RK=1,3,...,RKO
C         CI=SUMC(SGN(RK)*TM(RK))  IP=+1  RK=2,4,...,RKE
C                +EULER+XLOG
C         SHI=SUMOT(TM(RK))        IP=-1  RK=1,3,...,RKO
C         CHI=SUMET(TM(RK))        IP=+1  RK=2,4,...,RKE
C                +EULER+XLOG
C         EI=SUMOT+SUMET+EULER+XLOG                (X .GT. 0)
C               SGN(1)=1
C               SGN(RK+1)=-SGN(RK)          RK=1,3,...
C               SGN(RK+1)=+SGN(RK)          RK=2,4,...
C               TM(RK)=((T**RK)/(1*2...RK))/RK
C                     =PTM(RK)/RK
C                  PTM(1)=T
C                  PTM(RK+1)=PTM(RK)*(T/(RK+1))   RK .GE. 1
C               IF TM(RK)/SUM .LT. TOLER
C                 RKE=RK WHERE SUM=ABS(SUMC)        IC=1 OR 4
C                              SUM=SUMET            IC=2 OR 3
C                                          IC=4,X .GT. PSLSC
C                 RKO=RK WHERE SUM=ABS(SUMS)        IC=1 OR 4
C                              SUM=SUMOT            IC=2 OR 3
C                                          IC=4,X .GT. PSLSC
C         EXNEI= EI/EXP(T/2)/EXP(T/2)
C              =(EI/EXPHT)/EXPHT
C     CONTINUED FRACTION    T .GT. PSLSC
C         -CI+I(SI-PI/2)=E1(IT)
C                       =EXP(-IT)*(1 I/I (1+IT)-
C                                  1**2 I/I (3+IT)-
C                                  2**2 I/I (5+IT)-...)
C                       =EXP(-IT)*II(AM(RM) I/I BM(RM))
C                                            RM=1,2,...,RMF
C                            AM(1)=1
C                            AM(RM)=-(RM-1)**2     RM .GT. 1
C                            BM(RM)=2*RM-1+IT=BMR+I BMI
C                       =EXP(-IT)*(FM/GM)
C                       =EXP(-IT)*(FMR+I FMI)/(GMR+I GMI)
C                       =EXP(-IT)*F(RM)
C                       =(COST-I SINT)*(FR+I FI)
C         -CI+I(SI-PI/2)=(FR*COST+FI*SINT)+
C                                        I(FI*COST-FR*SINT)
C                         IF RESQ(RM) .LE. TOLSQ(=TOLER**2)
C                            OR RESQ(RM) .GE. RESQ(RM-1)
C                              (RESQ .GE. RESQP)
C                          RMF=RM  WHERE
C                            RESQ=(MOD(1-F(RM-1)/F(RM)))**2
C     ASYMPTOTIC EXPANSION    T .GT. AELL
```

```
117*      C          EI=(EXNEI*EXPHT)*EXPHT
118*      C          EXNEI=(1+SUME(TM(RK)))/T                RK=1,2,...,RKF
119*      C          SHI=CHI=EI/2
120*      C              TM(RK)=(1*2...RK)/(T**RK)
121*      C              TM(0)=1
122*      C              TM(RK)=(RK/T)*TM(RK-1)                    RK .GE. 1
123*      C                IF TM(RK) .LT. TOLER (CONVERGENCE) RKF=RK OR
124*      C                   TM(RK) .GE. TM(RK-1)(DIVERGENCE) RKF=RK-1
125*      C RANGE.
126*      C     FOR SI(X),CI(X), ABS(X) .LT. ULSC(UPPER LIMIT FOR
127*      C                                          SIN,COS ROUTINE)
128*      C          X=APPROXIMATELY 2**21, NBM=27
129*      C                          2**56, NBM=60
130*      C     FOR EXP(-X)*EI(X), X .LE. RINF
131*      C     FOR EI(X), X .LT. XMAXEI (APPROXIMATELY 92.5,  NBC=8,
132*      C                                             715.6,  NBC=11)
133*      C                  NBC=NUMBER OF BINARY DIGITS IN THE BIASED
134*      C                  CHARACTERISTIC OF A FLOATING POINT NUMBER
135*      C     FOR SHI(X),CHI(X), ABS(X) .LT. XMAXHF
136*      C         X=APPROXIMATELY 93.2,   NBC=8
137*      C                         716.3,   NBC=11
138*      C ACCURACY. THE MAXIMUM RELATIVE ERROR, EXCEPT FOR REGIONS
139*      C           IN THE IMMEDIATE NEIGHBORHOOD OF ZEROS,ON THE
140*      C           UNIVAC 1108 IS 4.5(-7) FOR SINGLE PRECISION COM-
141*      C           PUTATION AND 7.5(-17) FOR DOUBLE PRECISION COM-
142*      C           PUTATION.
143*      C PRECISION. VARIABLE - BY SETTING THE DESIRED VALUE OF NBM
144*      C                       OR A PREDETERMINED VALUE OF TOLER
145*      C MAXIMUM     UNIVAC 1108 TIME/SHARING EXECUTIVE SYSTEM
146*      C TIMING.      NBM=27   NBM=60
147*      C (SECONDS)     .0093    .070
148*      C STORAGE. 954 WORDS REQUIRED BY THE UNIVAC 1108 COMPILER
149*      C
150*      C
151*           SUBROUTINE SICIEI(IC,X,SI,CI,CII,EI,EXNEI,SHI,CHI,
152*          1                                              CHII,IERR)
153*      C          MACHINE DEPENDENT STATEMENTS
154*      C               TYPE STATEMENTS
155*           DOUBLE PRECISION X,SI,CI,CII,EI,EXNEI,SHI,CHI,CHII
156*           DOUBLE PRECISION A,AELL,AM,AMIN,ASUMSC,
157*          1        BMI,BMR,COST,EXPL,EXPHT,
158*          2        FI,FIP,FMI,FMM1I,FMM1R,FMM2I,FMM2R,FMR,FR,FRP,
159*          3        GMI,GMM1I,GMM1R,GMM2I,GMM2R,GMR,
160*          4        PSLL,PSLSC,PTM,RE,RESQ,RESQP,RK,RM,
161*          5        SCC,SFMI,SFMR,SGMI,SGMR,SGN,
162*          6        SINT,SUMC,SUME,SUME0,SUMET,SUMOT,SUMS,SUMSC,
163*          7        T,TEMP,TEMPA,TEMPB,TM,TMAX,TMM1,TOLER,TOLSQ,
164*          8        XLOG,XMAXEI,XMAXHF
165*           DOUBLE PRECISION RINF,ULSC,EULER,HALFPI,PI,ALOG2,
166*          1                 ZERO,ONE,TWO,FOUR
167*           DIMENSION A(4)
168*           EQUIVALENCE (FMR,A(1)), (FMI,A(2)), (GMR,A(3)),
169*          1            (GMI,A(4))
170*      C          CONSTANTS
171*           DATA EULER/.57721566490153286060D0/
172*           DATA HALFPI/1.57079632679489661900/
173*           DATA PI/3.14159265358979323900/
174*           DATA ALOG2/.69314718055994530940D0/
```

```
175*            DATA ZERO,ONE,TWO,FOUR /
176*           1     0.0D0,1.D0,2.D0,4.D0/
177*      C                 RINF=MAXIMUM MACHINE VALUE
178*      C                 ULSC=MAXIMUM ARGUMENT FOR SIN,COS ROUTINE
179*      C                        APPROX. 2**(NBM-6) OR 10**(S-2)
180*      C                                (S=SIGNIFICANT FIGURES)
181*      C                 NBM=ACCURACY DESIRED OR THE
182*      C                     MAXIMUM NUMBER OF BINARY DIGITS IN THE
183*      C                       MANTISSA OF A FLOATING POINT NUMBER
184*      C                 TOLER=UPPER LIMIT FOR RELATIVE ERRORS
185*      C                      =2**(-NBM)=APPROX. 10**(-S)
186*      C TOLER PRECOMPUTED MAY BE INSERTED IN A DATA STATEMENT AND
187*      C THE NBM DATA STATEMENT ELIMINATED
188*            DATA RINF/.898846567431157953AD308 /
189*            DATA ULSC/.7205759403792793GD17/
190*            DATA NBM / 60 /
191*            TOLER=TWO**(-NBM)
192*      C NOTE - ARGUMENT CHECKS PRECEDING FUNCTION REFERENCES
193*      C        NECESSITATE ADDITIONAL MACHINE DEPENDENT STATEMENTS
194*      C        IN THE STATEMENT NUMBER RANGE 140-150
195*      C          INITIALIZATION OF OUTPUT FUNCTIONS
196*            SI=RINF
197*            CI=RINF
198*            CII=RINF
199*            EI=ZERO
200*            EXNEI=RINF
201*            SHI=ZERO
202*            CHI=ZERO
203*            CHII=RINF
204*      C          VALIDITY CHECK ON INPUT PARAMETERS
205*      C               INDICATOR CHECK
206*      C                 SET IND=IC
207*      C                   CHANGE IND=4 IF IC .LT. 1 OR .GT. 4
208*            IND=IC
209*            IF (IND .LT. 1) GO TO 10
210*            IF (IND .GT. 4) GO TO 10
211*            GO TO 20
212*         10 IND=4
213*      C                 ARGUMENT CHECK
214*      C                   X .GE. 0    IERR=0
215*      C                   X .LT. 0    IERR=1
216*      C                               (ERROR RETURN IF IC=2)
217*         20 IERR=0
218*            T=X
219*         30 IF (T) 40,50,90
220*         40 T=-T
221*            IF (IND .EQ. 1) GO TO 30
222*            IERR=1
223*            IF (IND .NE. 2) GO TO 30
224*            IF (X .LT. ZERO) RETURN
225*      C          SPECIAL CASES
226*      C                 X=0
227*         50 IF (IND-2) 80,70,60
228*         60 SHI=ZERO
229*            CHI=-RINF
230*            CHII=ZERO
231*         70 EI=-RINF
232*            EXNEI=-RINF
```

```
233*            IF (IND .NE. 4) RETURN
234*         80 SI=ZERO
235*            CI=-RINF
236*            CII=ZERO
237*            RETURN
238*         90 IF (T .LT. ULSC) GO TO 140
239*      C                ABS(X) .GE. ULSC
240*            IF (IND-2) 130,110,100
241*        100 SHI=RINF
242*            CHI=RINF
243*            CHII=ZERO
244*            IF (IERR .EQ. 1) GO TO 120
245*        110 EI=RINF
246*            EXNEI=(ONE+(ONE/T))/T
247*        120 IF (IND .NE. 4) GO TO 1000
248*        130 SI=HALFPI
249*            CI=ZERO
250*            CII=ZERO
251*            GO TO 1000
252*      C          EVALUATIONS FOR ABS(X)(=T) .GT. 0 AND .LT. ULSC
253*      C               ADDITIONAL MACHINE DEPENDENT STATEMENTS
254*      C                    FUNCTION REFERENCES
255*      C                    CONTROL VARIABLES
256*        140 XLOG=DLOG(T)
257*            SINT=DSIN(T)
258*            COST=DCOS(T)
259*            EXPL =DLOG(RINF)
260*            XMAXEI=EXPL+DLOG(EXPL+DLOG(EXPL)) -ONE/EXPL
261*            XMAXHF=XMAXEI+ALOG2
262*            AELL=-DLOG(TOLER)
263*            AMIN=ONE/RINF
264*            PSLL=TWO*DSQRT(AMIN)
265*            PSLSC=TWO
266*      C                EXPONENTIAL FUNCTION DETERMINATION
267*            IF (T .LE. TOLER) GO TO 150
268*            IF (T .GE. XMAXHF) GO TO 160
269*            EXPHT=DEXP(T/TWO)
270*            GO TO 170
271*        150 EXPHT=ONE
272*            GO TO 170
273*        160 EXPHT=RINF
274*      C                METHOD SELECTION
275*        170 IF (T .LE. PSLSC) GO TO 200
276*            IF (IND .EQ. 1) GO TO 500
277*            IF (IND .EQ. 4) GO TO 500
278*        180 IF (T .GT. AELL) GO TO 800
279*            GO TO 230
280*      C                    INDICATOR TO COMPUTE EI,SHI,CHI
281*        190 IF (IND .EQ. 1) GO TO 1000
282*            IND=3
283*            GO TO 180
284*      C                    METHOD --- POWER SERIES
285*      C                      SI(X),CI(X),          T .LE. PSLSC
286*      C                      EI(X),SHI(X),CHI(X),  T .LE. AELL
287*      C                         LIMITING VALUES, T NEAR ZERO
288*        200 IF (T .GT. PSLL) GO TO 210
289*            SUMC=ZERO
290*            SUMET=ZERO
```

```
291*            SUMS=T
292*            SUMOT=T
293*            GO TO 360
294*      C                                INITIALIZATION FOR SI,CI
295*       210  IF (IND .NE. 1) GO TO 230
296*       220  SUMS=ZERO
297*            SUMC=ZERO
298*            SUMSC=ZERO
299*            SGN=ONE
300*            GO TO 240
301*      C                                INITIALIZATION FOR SHI,CHI(AND EI)
302*       230  SUMOT=ZERO
303*            SUMET=ZERO
304*            SUMEO=ZERO
305*            IF (IND .EQ. 4) GO TO 220
306*      C                                     IP -  INDICATOR FOR ODD OR
307*      C                                       EVEN TERMS
308*       240  IP=-1
309*            RK=ONE
310*            PTM=T
311*      C                                COMPUTATION OF (T**K)/(1*2...K)/K
312*       250  TM=PTM/RK
313*      C                                SUMMATION FOR SI(CI)
314*              IF (IND .NE. 1) GO TO 310
315*       260    SUMSC=SGN*TM+SUMSC
316*      C                                RELATIVE ERROR FOR SI(CI)
317*      C PARTIAL SUM OF ALTERNATING ODD(EVEN) TERMS MAY EQUAL ZERO
318*            ASUMSC=SUMSC
319*       270    IF (ASUMSC) 280,300,290
320*       280    ASUMSC=-ASUMSC
321*              GO TO 270
322*       290    RE=TM/ASUMSC
323*              GO TO 320
324*       300    RE=RINF
325*              GO TO 320
326*      C                                SUMMATION FOR SHI(CHI)(AND EI)
327*       310    SUMEO=TM+SUMEO
328*              IF (IND .EQ. 4) GO TO 260
329*      C                                RELATIVE ERROR FOR SHI(CHI)
330*              RE=TM/SUMEO
331*      C                                SIGN CHANGE AND SELECTION
332*      C                                OF SUMS OF ODD(EVEN) TERMS
333*       320    IF (IP .EQ. 1) GO TO 330
334*              SGN=-SGN
335*              SUMS=SUMSC
336*              SUMSC=SUMC
337*              SUMOT=SUMEO
338*              SUMEO=SUMET
339*              GO TO 340
340*       330    SUMC=SUMSC
341*              SUMSC=SUMS
342*              SUMET=SUMEO
343*              SUMEO=SUMOT
344*      C                                RELATIVE ERROR CHECK
345*       340    IF (RE .LT. TOLER) GO TO 360
346*      C                                ADDITIONAL TERMS
347*              RK=RK+ONE
348*      C                                     UNDERFLOW TEST
```

```
C UNDERFLOWS AFFECTING ACCURACY ARE AVOIDED. ALL OTHER
C UNDERFLOWS ARE ASSUMED TO BE SET EQUAL TO ZERO
        IF (T .GT. PSLSC) GO TO 350
        IF (PTM .LE. (AMIN*RK*RK)/T ) GO TO 360
 350    PTM=(T/RK)*PTM
        IP=-IP
        GO TO 250
C                              SI,CI EVALUATION
 360  IF (IND .NE. 1) GO TO 380
 370  SI=SUMS
      CI=(SUMC+XLOG)+EULER
      CII=ZERO
      GO TO 1000
C                              EI EVALUATION
 380  IF (X .LE. ZERO) GO TO 390
      EI=(SUMET+SUMOT+XLOG)+EULER
      EXNEI=(EI/EXPHT)/EXPHT
      IF (IND .EQ. 2) RETURN
C                              SHI,CHI EVALUATION
 390  SHI=SUMOT
      CHI=(EULER+SUMET)+XLOG
      CHII=ZERO
      IF (IND .NE. 4) GO TO 1000
      GO TO 370
C                       METHOD --- CONTINUED FRACTION
C                         SI(X),CI(X),       T .GT. PSLSC
C                         -CI(T) + I (SI(T)-HALFPI)=E1(IT)
C                            INITIALIZATION
 500  SCC=RINF/FOUR
      TOLSQ=TOLER*TOLER
      RM=ONE
      AM=ONE
      BMR=ONE
      BMI=T
      FMM2R=ONE
      FMM2I=ZERO
      GMM2R=ZERO
      GMM2I=ZERO
      FMM1R=ZERO
      FMM1I=ZERO
      GMM1R=ONE
      GMM1I=ZERO
      RESQP=RINF
      FRP=ZERO
      FIP=ZERO
C                              RECURRENCE RELATION
C                                FM=BM*FMM1 + AM*FMM2
C                                GM=BM*GMM1 + AM*GMM2
 510  FMR=BMR*FMM1R-BMI*FMM1I+AM*FMM2R
        FMI=BMI*FMM1R+BMR*FMM1I+AM*FMM2I
        GMR=BMR*GMM1R-BMI*GMM1I+AM*GMM2R
        GMI=BMI*GMM1R+BMR*GMM1I+AM*GMM2I
C                              CONVERGENT F=FM/GM
C                                TESTS TO AVOID INCORRECT RESULTS
C                                    DUE TO OVERFLOWS(UNDERFLOWS)
C                                  FINDING MAXIMUM(=TMAX) OF
C                                    ABSOLUTE OF FMR,GMR,FMI,GMI
C                                    FOR SCALING PURPOSES
```

```
407*                TMAX=ZERO
408*                I=1
409*         520    TEMP=A(I)
410*         530      IF (TEMP) 540,560,550
411*         540      TEMP=-TEMP
412*                  GO TO 530
413*         550      IF (TEMP .LE. TMAX) GO TO 560
414*                  TMAX=TEMP
415*         560      IF (I .GE. 4) GO TO 570
416*                  I=I+1
417*                  GO TO 520
418*         570    SFMR=FMR/TMAX
419*                SFMI=FMI/TMAX
420*                SGMR=GMR/TMAX
421*                SGMI=GMI/TMAX
422*                TEMP=SGMR*SGMR + SGMI*SGMI
423*                FR=(SFMR*SGMR+SFMI*SGMI)/TEMP
424*                FI=(SFMI*SGMR-SFMR*SGMI)/TEMP
425*      C                              RELATIVE ERROR CHECK
426*                TEMP=FR*FR+FI*FI
427*                TEMPA=(FRP*FR+FIP*FI)/TEMP
428*                TEMPB=(FIP*FR-FRP*FI)/TEMP
429*                TEMP=ONE-TEMPA
430*                RESQ =TEMP*TEMP+TEMPB*TEMPB
431*                IF (RESQ  .LE. TOLSQ) GO TO 590
432*                IF (RESQ .GE. RESQP) GO TO 580
433*      C                              ADDITIONAL CONVERGENTS
434*                AM=-RM*RM
435*                RM=RM+ONE
436*                BMR=BMR+TWO
437*                FMM2R=FMM1R
438*                FMM2I=FMM1I
439*                GMM2R=GMM1R
440*                GMM2I=GMM1I
441*                FMM1R=FMR
442*                FMM1I=FMI
443*                GMM1R=GMR
444*                GMM1I=GMI
445*                FRP=FR
446*                FIP=FI
447*                RESQP=RESQ
448*      C                              SCALING
449*      C SCALING SHOULD NOT BE DELETED AS THE VALUES OF FMR,FMI AND
450*      C GMR,GMI MAY OVERFLOW FOR SMALL VALUES OF T
451*                IF (TMAX .LT. SCC/(BMR-AM ) ) GO TO 510
452*                FMM2R=FMM2R/TMAX
453*                FMM2I=FMM2I/TMAX
454*                GMM2R=GMM2R/TMAX
455*                GMM2I=GMM2I/TMAX
456*                FMM1R=FMM1R/TMAX
457*                FMM1I=FMM1I/TMAX
458*                GMM1R=GMM1R/TMAX
459*                GMM1I=GMM1I/TMAX
460*                GO TO 510
461*      C                              DIVERGENCE OF RELATIVE ERROR
462*      C                                ACCEPT PRIOR CONVERGENT
463*       580   FR=FRP
464*             FI=FIP
```

```
C                              SI,CI EVALUATION
 590  SI=FI*COST-FR*SINT+HALFPI
      CI=-(FR*COST+FI*SINT)
      CII=ZERO
      GO TO 190
C                        METHOD --- ASYMPTOTIC EXPANSION
C                          EI(X),EXNEI(X)          X .GT. AELL
C                          SHI(T)=CHI(T)=EI(T)/2   T .GT. AELL
C                             INITIALIZATION
 800  IF (IND .NE. 2) GO TO 880
 810  SUME=ZERO
      RK=ZERO
      TM=ONE
C                              ADDITIONAL TERMS
 820  TMM1=TM
        RK=RK+ONE
        TM=(RK/T)*TM
C                              TOLERANCE CHECK
        IF (TM .LT. TOLER) GO TO 840
        IF (TM .GE. TMM1) GO TO 830
        SUME=SUME+TM
        GO TO 820
C                              DIVERGENT PATH
 830  SUME=SUME-TMM1
C                              EXNEI EVALUATION
 840  IF (X .LT. ZERO) GO TO 870
      EXNEI=(ONE+SUME)/T
C                              EI EVALUATION - X .LT. XMAXEI
      IF (T .GE. XMAXEI) GO TO 850
      EI=(EXNEI*EXPHT)*EXPHT
      GO TO 860
C                              EI - LIMITING VALUE, X .GE. XMAXEI
 850  EI=RINF
C                              SHI,CHI EVALUATION - T .LT. XMAXHF
 860  IF (IND .EQ. 2) RETURN
 870  IF (T .GE. XMAXHF) GO TO 1000
      SHI=((((  ONE+SUME)/T)/TWO)*EXPHT)*EXPHT
      CHI=SHI
      CHII=ZERO
      GO TO 1000
C                              SHI,CHI - LIMITING VALUE
C                                                    T .GE. XMAXHF
 880  IF ( T .LT. XMAXHF) GO TO 810
      SHI=RINF
      CHI=RINF
      CHII=ZERO
      IF ( X .GT. ZERO) GO TO 810
      GO TO 1010
C          ADJUSTMENTS FOR X .LT. 0
1000  IF (X .GT. ZERO) RETURN
1010  IF (IC .EQ. 3) GO TO 1020
      SI=-SI
      CII=-PI
      IF (IC .EQ. 1) RETURN
1020  SHI=-SHI
      CHII=-PI
      RETURN
      END
```

TABLE 1

X	SI(X)				CI(X)			
.0	.0				$-\infty$			
.1−001	.9999944444	6111108276	6470528517	85973−002	−.4027979520	9823920722	4397561453	54440+001
.2−001	.1999955556	0888852607	8665206276	10625−001	−.3334907338	8599613460	8102204879	74545+001
.3−001	.2999850004	0499380108	0675616964	14691−001	−.2929567223	9811175640	0020167602	13443+001
.4−001	.3999644461	5106467200	4469605623	17391−001	−.2642060133	3009493497	5931669690	80028+001
.5−001	.4999305607	6366745212	5334525894	59473−001	−.2419141543	5519082432	1138357349	19704+001
.6−001	.5998800129	5920656146	8561513856	83629−001	−.2237094916	8693029859	1937022506	78885+001
.7−001	.6998094724	5377692911	0744599342	97997−001	−.2083269121	9543103133	4741544991	04089+001
.8−001	.7997156101	6294499144	3834290267	56205−001	−.1950112552	8007321930	1863050699	50094+001
.9−001	.8995950984	0144401781	3419097919	40853−001	−.1832754260	4358445302	6218391310	18002+001
.1+000	.9994446110	8276950160	5921185541	90930−001	−.1727868386	6572966389	9772515290	65736+001
.2+000	.1995560885	2623382140	0456944764	16595+000	−.1042205595	6727819753	6291690667	58895+001
.3+000	.2985040438	0704316138	6446229574	64345+000	−.6491729329	7116174495	6181105135	48965+000
.4+000	.3964614647	5137288302	0334263135	17445+000	−.3788093464	2524433208	4661757229	44184+000
.5+000	.4931074180	4306668916	1626707572	76465+000	−.1777840788	0661290133	5810271070	56908+000
.6+000	.5881288096	0808006689	9647904006	83682+000	−.2227070695	9279762526	6027439020	62561−001
.7+000	.6812222391	1661131088	9506811453	94252+000	.1005147070	0889783268	9135892810	28616+000
.8+000	.7720957854	8199656025	3889712479	89549+000	.1982786159	5246717701	5833582233	84097+000
.9+000	.8604707107	4529293277	4085411696	29011+000	.2760678304	6777286015	0434115538	80334+000
.1+001	.9460830703	6718301494	1353313823	17966+000	.3374039229	0096813466	2646203889	15077+000
.2+001	.1605412976	8026948485	7672014819	85889+001	.4229808287	7486499569	8565153198	25589+000
.3+001	.1848652527	9994682563	9773025111	19732+001	.1196297860	0800032762	6472281176	67785+000
.4+001	.1758203138	9490530581	0555930335	85016+001	−.1409816978	8693041163	9144898694	03593+000
.5+001	.1549931244	9446741372	7440840073	06390+001	−.1900297496	5664387861	8458900116	30081+000
.6+001	.1424687551	2805065357	6903102791	71420+001	−.6805724389	3247126204	1683048406	17428−001
.7+001	.1454596614	2480935906	1476849383	61604+001	.7669527848	2184518382	9157630314	36847−001
.8+001	.1574186821	7069420520	8297145120	66585+001	.1224338825	3200955729	2295958268	73503+000
.9+001	.1665040075	8296024951	0665342789	71085+001	.5534753133	3133607085	6416484497	93252−001
.1+002	.1658347594	2188740493	3097187938	96725+001	−.4545643300	4455372634	5328299526	27853−001
.2+002	.1548241701	0434398401	6364334212	95137+001	.4441982084	5353316539	7687169925	70578−001
.3+002	.1566756540	0303511109	8373130900	67982+001	−.3303241728	2071143779	2264409630	03714−001
.4+002	.1586985119	3547845067	7566596201	46420+001	.1902000789	6208766961	9812034202	67962−001
.5+002	.1551617072	4859358947	2798559485	93775+001	−.5628386324	1163054401	8589549846	40996−002
.6+002	.1586745616	2599474123	2644013231	99104+001	−.4813243377	4432152888	7234385245	11961−002
.7+002	.1561594849	1780061055	2298220467	79853+001	.1092198847	3464977037	8113544132	70120−001
.8+002	.1572330886	9124873153	5125172966	74798+001	−.1240250115	5070958192	0458829452	17856−001
.9+002	.1575663406	6574562607	3805334080	46545+001	.9986124071	6431336804	5713422489	45291−002
.1+003	.1562225466	8890562933	5234513880	45027+001	−.5148825142	6104921444	4355390534	44979−002
.2+003	.1568382339	3394698333	5878557542	35465+001	−.4378446093	0278256791	6569771749	32552−002
.3+003	.1570881088	2137495192	5231225344	08620+001	−.3332199918	5921117799	7045482583	87851−002
.4+003	.1572114869	2738117518	0132144796	40848+001	−.2123988830	8463489343	4793666984	71878−002
.5+003	.1572565882	2431687035	3434162096	10243+001	−.9320008144	0429025451	7261962053	86680−003
.6+003	.1572461233	9493979398	3169426317	07478+001	.7641202377	3809449584	1487520193	67238−004
.7+003	.1571993932	2374915706	3702809228	14464+001	.7788100127	3975633981	0132017166	41226−003
.8+003	.1571355087	6214727479	0846382718	80577+001	.1118158760	9180513118	2640905181	04941−002
.9+003	.1570721487	6829785964	0335292159	62388+001	.1108585782	0500159240	9398971297	71935−002
.1+004	.1570233121	9687712181	4796277803	63344+001	.8263155110	9068228200	1773882343	20723−003

$\pi/2$ = 1.57079 63267 94896 61923 13216 91639 75144

TABLE 2

X	EI(X)				EXP(−X)*EI(X)			
.0	−∞				−∞			
.1−001	−.4017929465	4266693867	7534341058	83082+001	−.3977950399	2615576971	3408360703	19476+001
.2−001	−.3314706894	4101539023	9227816770	05873+001	−.3249071300	3015878880	1888549128	04425+001
.3−001	−.2899115723	9402794440	9618312164	59357+001	−.2813433905	5380940257	4977395174	78007+001
.4−001	−.2601256577	5728261871	6902308916	05352+001	−.2499259848	2574874044	8538470778	63194+001
.5−001	−.2367884598	5793745242	5950757147	49992+001	−.2252401503	9907626374	1216434311	86807+001
.6−001	−.2175282915	5516236358	6147208749	96411+001	−.2048604300	3782595071	0809271777	33217+001
.7−001	−.2010800063	5428875184	2790326756	89313+001	−.1874857552	3138763197	4860337705	30829+001
.8−001	−.1866884102	7729969797	4868911203	84179+001	−.1723351232	0791016360	9393503175	42803+001
.9−001	−.1738663750	3469511928	7135014650	31175+001	−.1589019022	1427078812	0803355460	85013+001
.1+000	−.1622812813	9692766749	6568299922	74753+001	−.1468381756	5476302940	1938294633	01893+001
.2+000	−.8217605879	0240031565	3310869899	97772+000	−.6728006649	8313731896	8600992150	20087+000
.3+000	−.3026685392	6582588446	8136901536	93942+000	−.2242223687	1524378927	5487443901	40039+000
.4+000	.1047652186	1932479322	8929763883	46026+000	.7022622616	7839611020	0870681109	89467−001
.5+000	.4542199048	6317357992	0523812662	80237+000	.2754982985	5127026213	2219977961	61145+000
.6+000	.7698812899	3735943709	3137330257	85125+000	.4225198103	2870176143	1329955834	52537+000
.7+000	.1064907194	6242905405	7162359274	67593+001	.5288172627	5216097535	3053374545	58451+000
.8+000	.1347396548	2123259381	1899638330	35829+001	.6054242952	6336443347	4212274854	28025+000
.9+000	.1622811713	6968674413	4419142440	54222+001	.6597860062	6079393891	3476848207	67092+000
.1+001	.1895117816	3559367554	6652093433	16343+001	.6971748832	3506606876	5478681919	55160+000
.2+001	.4954234356	0018901633	7950513022	70353+001	.6704827097	9007328104	3223808083	90335+000
.3+001	.9933832570	6254165580	0833601921	67653+001	.4945764013	4864123502	8769696756	17092+000
.4+001	.1963087447	0056220022	6457202797	23839+002	.3595520078	6362069617	7421675826	60745+000
.5+001	.4018527535	5803177455	0914217937	95867+002	.2707662554	9105719558	0378348094	65324+000
.6+001	.8598976214	2439204803	5834003079	90690+002	.2131473100	8159360315	3856832629	71120+000
.7+001	.1915047433	3550139595	3063148272	45695+003	.1746297217	6579015129	0938352318	25166+000
.8+001	.4403798995	3483826899	7424596659	39339+003	.1477309983	7340099664	7174203305	76383+000
.9+001	.1037878290	7170895876	5757322679	36222+004	.1280843565	2321386813	5996461057	99933+000
.1+002	.2492228976	2418777591	3844014399	85248+004	.1131470204	7341077803	4051681354	47701+000
.2+002	.2561565266	4056588820	4811208040	98072+008	.5279779527	9648132254	5779817983	87276−001
.3+002	.3689732094	0727419706	4006328910	84575+012	.3452712179	2361846131	5818701748	98032−001
.4+002	.6039718263	6112415783	5923141851	06913+016	.2565886278	5975145205	8133890489	43081−001
.5+002	.1058563689	7131690963	0615414332	29987+021	.2041704555	5943987333	5659940616	10444−001
.6+002	.1936182213	9292765388	2072596687	52373+025	.1695420039	4813286610	0796814766	65768−001
.7+002	.3646352759	5797356367	1345154856	77168+029	.1449589211	5101025543	4076487427	26111−001
.8+002	.7014600004	9047999696	2996948047	97919+033	.1266031055	4032884299	9914561805	60367−001
.9+002	.1371416869	5072519995	0222595544	66918+038	.1123740714	1495976244	6676886934	83357−001
.1+003	.2715552744	8538798219	1401464231	08254+042	.1010206252	7748357112	3003599184	99944−001
.2+003	.3631235233	1593568523	9671004384	64250+085	.5025253826	9333012313	0841542901	12999−002
.3+003	.6496482508	0886657890	2569189493	42473+128	.3344519269	3037826333	1099042135	67540−002
.4+003	.1308647281	7074277342	4939487839	11718+172	.2506281486	7484941759	1294495095	34843−002
.5+003	.2812821397	8862943374	7493151789	64387+215	.2004016096	7757734712	6302843056	26131−002
.6+003	.6298882891	3879314245	2569543511	94621+258	.1669453750	3112400160	6469063742	31301−002
.7+003	.1450978736	0525608526	2088252210	93011+302	.1430618100	9351634011	0181288435	03898−002
.8+003	.3412238865	4483770461	9667674034	14661+345	.1251566420	9721409147	4398485683	85823−002
.9+003	.8152195006	2752522682	3560124386	87339+388	.1112348431	6823895903	3663223603	11253−002
.1+004	.1972045137	1412383028	0964504841	20236+432	.1001002006	0241207250	8068654920	21171−002

TABLE 3

X	SHI(X)				CHI(X)			
.0	.0				−∞			
.1−001	.1000005555	5722222505	6692404330	26380−001	−.4027929520	9823916092	8101265102	13346+001
.2−001	.2000044444	9777814059	1136870933	39861−001	−.3334707338	8599317164	5139185479	39271+001
.3−001	.3000150004	0500619903	9860097325	50634−001	−.2929117223	9807800640	0016913137	84864+001
.4−001	.4000355572	6226866293	4193925798	75059−001	−.2641260133	2990530534	6244248174	04103+001
.5−001	.5000694496	5299922650	1974168087	39980−001	−.2417891543	5446744469	0970498828	37392+001
.6−001	.6001200129	6079350024	5724232696	91766−001	−.2235294916	8477029858	8604451076	93328+001
.7−001	.7001905835	6955665134	1973404093	50990−001	−.2080819121	8998431835	6210060797	82823+001
.8−001	.8002844990	6372249715	1895071106	72843−001	−.1946912552	6793692294	6387861914	90908+001
.9−001	.9004050984	2855835469	1051467784	84646−001	−.1828704260	1898070283	4045529328	16021+001
.1+000	.1000555722	2505699555	7615329532	17784+000	−.1722868386	1943336705	2329832875	96531+001
.2+000	.2004449781	4074638634	0730853837	22252+000	−.1022205566	0431467019	9404172373	72002+001
.3+000	.3015040562	0501041398	1095310302	38247+000	−.6041725954	7083629844	9232211839	32189+000
.4+000	.4035726687	4249363590	5979378947	55253+000	−.2988074501	2316884267	7049615064	09227+000
.5+000	.5069967498	1966719583	3659875988	94380+000	−.5277684495	6493615913	1360633261	41435−001
.6+000	.6121303965	6338077262	4562784597	54146+000	.1577508933	7397866446	8574545660	30978+000
.7+000	.7193380189	2889984241	9121259127	50575+000	.3455691756	9539069815	2502333619	25356+000
.8+000	.8289965633	7893448638	6910469189	60092+000	.5183999848	3339145173	2085914113	98201+000
.9+000	.9414978265	1143354092	2701645733	42970+000	.6813138871	8543390042	1489778671	99251+000
.1+001	.1057250875	3757285145	7184235489	58780+001	.8378669409	8020824089	4678579435	75631+000
.2+001	.2501567433	3549756414	7337248272	75424+001	.2452666922	6469145219	0613264749	94929+001
.3+001	.4973440475	8598067977	1041838252	27051+001	.4960392094	7656097602	9791763669	40601+001
.4+001	.9817326911	2330344645	6229756992	81526+001	.9813547558	8231855580	8342270979	56862+001
.5+001	.2009321182	5697226390	4443761778	82843+002	.2009206353	0105951064	6470456159	13024+002
.6+001	.4299506111	2445683731	1213478510	53231+002	.4299470102	9993521072	4620524569	37459+002
.7+001	.9575242940	8616503145	6397896419	56496+002	.9575231392	6884892807	4233586305	00452+002
.8+001	.2201899686	0023055646	1163184608	69467+003	.2201899309	3460771253	6261412050	69872+003
.9+001	.5189391515	8222188283	1922673971	09373+003	.5189391391	3486770482	5650552822	52849+003
.1+002	.1246114490	1994233444	1188221070	06923+004	.1246114486	0424544147	2655793329	78325+004
.2+002	.1280782633	2028294459	4181868552	98444+008	.1280782633	2028294361	0629339487	99627+008
.3+002	.1844866047	0363709853	2003165966	19888+012	.1844866047	0363709853	2003162944	64687+012
.4+002	.3019859131	8056207891	7961570925	53457+016	.3019859131	8056207891	7961570925	53456+016
.5+002	.5292818448	5658454815	3077071661	49936+020	.5292818448	5658454815	3077071661	49936+020
.6+002	.9680911069	6463826941	0362983437	61866+024	.9680911069	6463826941	0362983437	61866+024
.7+002	.1823176379	7898678183	5672577428	38584+029	.1823176379	7898678183	5672577428	38584+029
.8+002	.3507300002	4523999848	1498474023	98960+033	.3507300002	4523999848	1498474023	98960+033
.9+002	.6857084347	5362599975	1112977723	34589+037	.6857084347	5362599975	1112977723	34589+037
.1+003	.1357776372	4269399109	5700732115	54127+042	.1357776372	4269399109	5700732115	54127+042
.2+003	.1815617616	5796784261	9835502192	32125+085	.1815617616	5796784261	9835502192	32125+085
.3+003	.3248241254	0443328945	1284594746	71236+128	.3248241254	0443328945	1284594746	71236+128
.4+003	.6543236408	5371386712	4697439195	58592+171	.6543236408	5371386712	4697439195	58592+171
.5+003	.1406410698	9431471687	3746575894	82193+215	.1406410698	9431471687	3746575894	82193+215
.6+003	.3149441445	6939657122	6284771755	97311+258	.3149441445	6939657122	6284771755	97311+258
.7+003	.7254893680	2628042631	0441261054	65053+301	.7254893680	2628042631	0441261054	65053+301
.8+003	.1706119432	7241885230	9833837017	07331+345	.1706119432	7241885230	9833837017	07331+345
.9+003	.4076097503	1376261341	1780062193	43669+388	.4076097503	1376261341	1780062193	43669+388
.1+004	.9860225685	7061915140	4822524206	01178+431	.9860225685	7061915140	4822524206	01178+431

JOURNAL OF RESEARCH of the National Bureau of Standards—B. Mathematical Sciences
Vol. 80B, No. 2, April–June 1976

Definite Integrals of the Complete Elliptic Integral *K**

M. L. Glasser**

(January 26, 1976)

Definite integrals whose integrands contain the complete elliptic integral of the first kind are compiled.

Key words: Complete elliptic integral; cubic lattice; definite integral; special functions.

INTRODUCTION

In studying a wide variety of problems involving three dimensional lattices, one is led to examine integrals of the form

$$\int\int\int_0^{2\pi} F(\cos xx,\ \cos my,\ \cos nz)dxdydz. \tag{1}$$

In many cases two of the integrations can be carried out explicitly yielding an integral of the form

$$\int_\alpha^\beta F(u)\ \mathbf{K}(u)du \tag{2}$$

where

$$\mathbf{K}(u) = \int_0^{\pi/2} [1 - u^2 \sin^2\theta]^{-1/2}\, d\theta \tag{3}$$

is the complete elliptic integral of the first kind. Examples of this are contained in references [1][1] and [2]. In the studying of the vibrational properties of a harmonically coupled simple cubic lattice of mass points, the author was led to evaluate a large number of these integrals.

*An invited paper. This work was supported by the National Research Council of Canada under Grant No. A9344.
**Present address: Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

[1] Figures in brackets indicate the literature references at the end of this paper.

The literature concerning integrals of the form (2) is rather sparse. The known results, due mainly to Kaplan [3] and Byrd and Friedman [4] are summarized by Gradshteyn and Ryzhik [5] as entries 6.141–6.147 and 5.153. For completeness these older results have been included in the following table. The methods by which these formulas have been obtained are diverse and it would be impractical to go into the derivations in detail. Several, such as (1–18) were obtained quite simply by integrating the series representing $\mathbf{K}(u)$ term by term. Others, e.g., (II–10) were found by applying the Fourier inversion theorem to a known integral. Several results were obtained by writing $\mathbf{K}(u)$ as a hypergeometric function and applying the general formulas of section 7.5 of reference [5]. Occasionally more complicated procedures were needed (e.g., see [6]). When possible the results were independently checked by considering special cases.

It is felt that the publication of a table of these formulas would be useful to others working in this field as well as form a nucleus for a more extensive tabulation by others in the future. In addition, a number of our results have a certain intrinsic charm.

The notation used in the following tables is that of the Bateman manuscript project series [7]. In addition, $\theta(u)$ denotes the unit step function which is unity when $u > 0$ and vanishes for $u < 0$. The conditions on the parameters for which the formulas are valid, when not explicitly noted, are easily ascertained and are generally those under which both sides are defined. Finally, these table partially supplement the recent tabulation of representations of the complete elliptic integral as Bessel function integrals by Okui [8].

I. *Algebraic integrands*

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)du$
(1)	$u^{s-1}(1-u^2)^{(1/2)-s}$	$\frac{\pi^{1/2}}{8}\frac{\Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)}\frac{\Gamma\left(\frac{1}{2}s\right)}{\Gamma\left(1-\frac{1}{2}s\right)}\frac{\Gamma\left(\frac{3}{2}-s\right)\Gamma\left(\frac{3}{4}-\frac{1}{2}s\right)}{\Gamma\left(\frac{5}{4}-\frac{1}{2}s\right)}\left(0<\operatorname{Re}s<\frac{3}{2}\right)$ (1–18)
(2)	$u^{s-1}(1-u^2)^{(1/2)s-1}$	$\frac{2^{-s}\pi^2}{2}\frac{\Gamma^2\left(\frac{1}{2}s\right)}{\Gamma^2\left(\frac{3}{4}\right)\Gamma^2\left(\frac{1}{2}s+\frac{1}{4}\right)}\quad(0<\operatorname{Re}s)$ (1–18)
(3)	$u^{s-1}(1-u^2)^{-s/2}$	$\frac{\pi^{5/2}}{4}\csc\left(\frac{1}{2}\pi s\right)\frac{\Gamma\left(\frac{3}{4}+\frac{1}{4}s\right)\Gamma\left(\frac{3}{4}-\frac{1}{4}s\right)}{\Gamma^2\left(\frac{3}{4}\right)\Gamma\left(\frac{1}{2}+\frac{1}{4}\right)\Gamma\left(1-\frac{1}{4}s\right)}\quad(0<\operatorname{Re}s<2)$ (1–18)
(4)	$u^{s-1}(1-u^2)^{-(s+1)/4}$	$\frac{\pi^{3/2}}{4}\frac{\Gamma\left(\frac{1}{2}s\right)\Gamma^2\left(\frac{3}{4}-\frac{1}{4}s\right)}{\Gamma^2\left(\frac{3}{4}\right)\Gamma\left(\frac{1}{4}s+\frac{1}{2}\right)\Gamma\left(1-\frac{1}{4}s\right)}\quad(0<\operatorname{Re}s<3)$ (1–18)
(5)	$u(u^2+z^2)^{-1}+(1+u^2z^2)^{-1}$	$\frac{\pi}{2}(1+z^2)^{-1/2}\mathbf{K}[(1+z^2)^{-1/2}]$
(6)	$\theta(u-z)(1-u^2)^{-1/2}(u^2-z^2)^{-1/2}$	$2(1+z)^{-1}\mathbf{K}\left[\left(\frac{1-z}{1+z}\right)^{1/2}\right]\mathbf{K}\left[\left(\frac{2z}{1+z}\right)^{1/2}\right]$ $(0<z<1)$

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)du$	
(7)	$\int_0^1 (1-u^2)^{-1/2}\mathbf{K}[2(z^2u^2+z(1-z))]du = \mathbf{K}(k_+)\mathbf{K}(k_-)$		
	$k_\pm^2 = [1-(1-4z)^{1/2}]^4 + z[1 \pm (1-z)^{1/2}]^2, \quad \operatorname{Re} z < \frac{1}{4}$		
(8)	$u(1-u^2)^{-1/2}/(z^2u^2+1)$	$\frac{\pi}{2}(z^2+1)^{-1}\mathbf{K}[z/(z^2+1)^{1/2}]$	
		$z > 0$	
(9)	$u(1-u^2)^{-1/2}/(1-z^2u)$	$\frac{\pi}{2}(1-z^2)^{-1/2}\mathbf{K}(z)$	
		$0 \le z < 1$	
(10)	$\theta(z-u)(1-u^2)^{-1/2}(z^2-u^2)^{-1/2}$	$(z+1)^{-1}\,\mathbf{K}^2[(2z)^{1/2}/(1+z)^{1/2}]$	
		$0 \le z < 1$	(1–17)
(11)	$u(1-u^2)^{-1/2}(z^2-u^2)^{-1/2}$	$(1+z)^{-1}\mathbf{K}^2[(2/(1+z))^{1/2}]$	
		$z > 1$	
(12)	$\theta(\sin\phi-u)u$	$\mathbf{E}(\sin\phi) - \cos^2\phi\ \mathbf{K}(\sin\phi)$	
		$(0 \le \phi \le \pi)$	(1–18)
(13)	$\theta(u-z)u^{-2}$	$z^{-1}\,\mathbf{E}(z) - 1$	
		$(z > 0)$	(1–18)
(14)	$(1-u^2)^{\nu-1}u^{-2\nu}$	$2^{4\nu-4}\sec(\pi\nu)\,\frac{\Gamma^4(\nu)}{\Gamma^2(2\nu)}$	
		$0 < \nu < \frac{1}{2}$	(1–18)
(15)	$(1+u)^{-1}$	$\pi^2/8$	[5]
(16)	$\int_0^1 \mathbf{K}[2^{-1/2}u]u(1-u^2)^{\nu-1}\,du = \frac{\pi^{3/2}\,\Gamma(\nu)}{2^{\nu+2}\,\Gamma^2\left(\frac{1}{2}\nu+\frac{3}{4}\right)}$		
	$(\operatorname{Re}\nu > 0)$		(1–18)
(17)	$\theta(z-u)(z^2-u^2)^{\nu-1}\,(1u^2)^{\frac{1}{2}-\nu-\sigma}u^{2\nu-1}$	$\frac{\pi}{4}B(\sigma,\nu)\frac{z^{2(\sigma+\nu-1)}}{(1-z^2)^\sigma}\ {}_3F_2\left((1/2),(1/2),\sigma;1,\sigma+p-\frac{z^2}{1-z^2}\right)$	
		$\nu,\sigma > 0,\ z^2 < 1$	
(18)	$\int_0^1 \mathbf{K}(zu)u^{2\nu-1}\,(1-u^2)^{\sigma-1}du = \frac{\pi}{4}\frac{\Gamma(\nu)\Gamma(\sigma)}{\Gamma(\nu+\sigma)}\ {}_3F_2\left(\frac{1}{2},\frac{1}{2},\nu;1,\sigma+\nu;z^2\right)$		
	$\nu,\sigma > 0,\ \lvert\arg(1-z^2)\rvert < \pi$		
(19)	$(1-u^2)^{-1/2}\,\frac{u[(1-z^2)-u^2]}{[(1+z^2)-u^2]^2}$	$\frac{\pi}{4}z^{-2}\ {}_2F_1\left(\frac{1}{2},\frac{3}{2};1;-z^{-2}\right)$	
(20)	$\int_0^1 \mathbf{K}(2^{-1/2}u)u^{2\sigma-1}\,(1-u^2)^{\nu-1}\,(2-u^2)^{\sigma}\,du = \frac{\pi}{8}\frac{\Gamma(\sigma+1)\Gamma\left(\frac{1}{2}\nu\right)}{\Gamma\left(1+\sigma+\frac{1}{2}\nu\right)}\ {}_3F_2\left(\frac{1}{4},\frac{1}{4},1+\sigma;1,1+\frac{1}{2}\nu+\sigma;1\right)$		
	$(\operatorname{Re}\nu > 0,\ \operatorname{Re}\sigma > -1)$		

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)du$
(21)	$u(1-u^2)^{\nu-1}(1-zu^2)^{-\sigma}$	$\frac{\pi}{4}\frac{\Gamma^2(\nu)}{\Gamma^2\left(\nu+\frac{1}{2}\right)}\ (1-z)^{\sigma}\ {}_3F_2\left(\nu,\nu,\sigma;\nu+\frac{1}{2},\nu+\frac{1}{2};\frac{z}{z-1}\right)$ $\nu>0,\ \lvert\arg(1-z)\rvert<\pi$
(22)	$\int_0^1(1-u^2)^{-1/2}\mathbf{K}(au)du=\frac{\pi^2}{4}\left\{{}_2F_1\left(\frac{1}{4},\frac{1}{4};1;a^2\right)\right\}^2$ $\operatorname{Re} a\geq 0$ (1–18)	
(23)	$\theta(z-u)u(z^2-u^2)^{\nu-1}$	$\frac{\pi}{4}\Gamma(\nu)z^{\nu}(1-z^2)^{\nu/2}P^{-\nu}_{-(1/2)}(1-2z^2)$ $0<z<1,\quad \nu>0$
(24)	$\theta(u-z)u(1-u^2)^{\nu}(u^2-z^2)^{-\nu-(3/2)}$	$\frac{2^{-(3/2)}}{(1-z^2)^{\nu+(3/2)}}\frac{\Gamma^2(\nu+1)\Gamma\left(-\frac{1}{2}-\nu\right)}{(c^2-1)^{2\nu+1)/4}}P^{-\nu-(1/2)}_{-1/2}(c)$ $c=(1+z)^2/(1-z^2),\ 0\leq z<1,\ -1<\operatorname{Re}\nu<-\frac{1}{2}$ (1–17)
(25)	$\theta(z-u)u(z^2-u^2)^{\nu-1}/(1-u^2)^{\nu+(1/2)}$	$\frac{\pi}{4}\Gamma(\nu)z^{\nu}(1-z^2)^{-1}P^{-\nu}_{-1/2}[(1+z^2)/(1-z^2)]$ $\nu>0,\quad 0<z<1$
(26)	$u^{-1}(1-u^2)^{-1/2}\left[1+\frac{z}{(z^2+x^2)^{1/2}}\right]$ $x^2=(1-u^2)/u^2$	$(C/2)^{1/2}\frac{4}{\pi}\left[\left(\frac{2}{\xi+1}\right)(\xi-\sqrt{\xi^2-1})\right]^{1/2}$ $\mathbf{K}\left(\sqrt{\frac{\xi-1}{\xi+1}}\right)\mathbf{K}(\xi-\sqrt{\xi^2-1}),\ \xi=\sqrt{Cz}$ $z>1,\ C=2[z-(z^2-1)^{1/2}]$
(27)	$\theta(a-u)(a^2u^{-2}-1)^{\nu-1}u^{-2}$	$\frac{1}{4}e^{\nu\pi i}(c^2-1)^{\nu/2}(c+1)^{1-\nu}Q^{-\nu}_{-(1/2)}(c)$ $c^2=2a^{-2}-1,\ 0<a<1,\ 0<\operatorname{Re}\nu<\frac{1}{2}$
(28)	$\theta[(1+z^2)^{-1/2}-u][u(1-u^2)]^{-1/2}[(1-u^2)^{1/2}-zu]^{-1/2}$	$\frac{1}{2}z^{-1/2}[Q_{-3/4}(z+z^{-1})P_{-3/4}(z+z^{-1})$ $-Q_{-1/4}(z+z^{-1})P_{-1/4}(z+z^{-1})]$ $(z>0)$
(29)	$\theta(z-u)(1-u^2)^{-3/4}\left[\frac{z(1-u^2)+u(1-z^2)}{z^2-u^2}\right]^{1/2}$	$\left(\frac{1-z^2}{z^2}\right)^{1/4}\left\{[Q_{-1/4}(C)]^2+[Q_{-3/4}(C)]^2\right\}$ $0<z<1,\ C^{-1}=z(1-z^2)^{1/2}$
(30)	$\theta[(1+z^2)^{-(1/2)}-u]u^2\left[\frac{u^2}{(1-u^2)-z^2u^2}\right]^{\nu+(3/2)}$	$\frac{2^{(6\nu+11)/4}}{\lvert z\rvert^{\nu+1/2}}e^{-(\nu+1/2)\pi i}Q^{\nu+(1/2)}_{-(1/2)}(2z^2+1)$ z real, $-3<\operatorname{Re}\nu<-\frac{1}{2}$
(31)	$(1-u^2)^{\nu-1}\{u+z(1-u^2)^{1/2}]^{-2\nu}$ $+\lvert u-z(1-u^2)^{1/2}\rvert^{-2\nu}\}$	$2^{4\nu-3}\frac{\Gamma^4(\nu)}{\Gamma^2(2\nu)}\sec(\pi\nu)\times$ $\times\ {}_3F_2(\nu,\nu,\nu;\nu+1/2,1/2;-z^2)$ $\left(0<z<1,\ \lvert\nu\rvert<\frac{1}{2}\right)$

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)\,du$
(32)	$(1-u^2)^{\nu-1}\left\{[u+z(1-u^2)^{1/2}]^{-2\nu} + \dfrac{\operatorname{sgn}[z(1-u^2)^{1/2}-u]}{\lvert z(1-u^2)^{1/2}-u\rvert^{2\nu}}\right\}$	$\dfrac{4^{2\nu}}{\pi}\dfrac{z\,\Gamma(1-2\nu)}{\Gamma(1+2\nu)}\Gamma^4\left(\dfrac{1}{2}+\nu\right)\cos\pi\nu\ {}_3F_2[\nu+(1/2),\nu+(1/2),\nu+(1/2);(3/2),\nu+1;-z^2]$ $\left(0<z<1)\,(\lvert\nu\rvert<\dfrac{1}{2}\right)$
(33)	$\theta[u-z(1+z^2)^{-1/2}]u(1-u^2)^{-3/2}\cdot\left[\dfrac{1-u^2}{u^2-(1-u^2)z^2}\right]^{(2\nu+3)/2}$	$\dfrac{\Gamma^2(\nu+1)\Gamma\left(-\nu-\dfrac{1}{2}\right)2^{(\nu-3)/2}}{2^{(1/4)}\,z^{\nu+(1/2)}}P_{-1/2}^{-\nu-1/2}(2z^2+1)$ $\left(-1<\nu<-\dfrac{1}{2}\right)$, z real
(34)	$u^{-1/2}(1-u^2)^{-1/2}\left\{[zu+(1-u^2)^{1/2}]^{-1/2} + \dfrac{\operatorname{sgn}[zu-(1-u^2)^{1/2}]}{\lvert zu-(1-u^2)^{1/2}\rvert^{1/2}}\right\}$	$z^{-1/2}Q_{-(1/4)}(z+z^{-1})P_{-(1/4)}(z+z^{-1})$ z real
(35)	$(1-u^2)^{-3/4}\left\{[u+z(1-u^2)^{1/2}]^{-1/2} + \dfrac{\operatorname{sgn}[z(1-u^2)^{1/2}-u]}{\lvert z(1-u^2)^{1/2}-u\rvert^{1/2}}\right\}$	$z^{-1/2}[Q_{-1/4}(z+z^{-1})]^2$ z real
(36)	$(1-u^2)^{-3/4}\{\lvert z(1-u^2)^{1/2}-u\rvert^{-1/2} + \lvert z(1-u^2)^{1/2}+u\rvert^{-1/2}\}$	$z^{-1/2}\{Q_{-3/4}(z+z^{-1})\}^2$ z real
(37)	$u^{-1/2}(1-u^2)^{-1/2}\{[zu+(1-u^2)^{1/2}]^{-1/2} + \lvert zu-(1-u^2)^{1/2}\rvert^{-1/2}\}$	$z^{-1/2}Q_{-3/4}(z+z^{-1})P_{-3/4}(z+z^{-1})$ z real
(38)		$\int_1^z (u^2-1)^{-1/2}(z^2-u^2)^{-1/2}\mathbf{K}(au)\,du = \dfrac{2}{z[1+(1-a^2)^{1/2}]}\mathbf{K}(k_+)\mathbf{K}'(k_-)$ $k_\pm = \dfrac{1\pm(1-a^2z^2)^{1/2}}{z[1+(1-a^2)^{1/2}]}$, z real
(39)		$\int_0^1 k_+^{-1/2}\,\mathbf{K}(k_-)du = \dfrac{\pi^2\phi}{4}\csc\left(\dfrac{1}{2}\phi\right)$ $k_\pm = (1-u^2)\cos\phi + u^2 \pm [(1-u^2)^2\cos^2\phi + u^4 + 2u^2(1-u^2)\cos\phi - 1]^{1/2}$ $0\le\phi\le\pi$
(40)	$u^{-1}(1-u^2)^{-1/2}(2-u^2)^{-1}\ln(1-u^2)$	$-2(\pi/2)^{3/2}\operatorname{arcsinh}(1)$
(41)	$[u/(1-u^2)]^{1/2}\ln\lvert(u^{1/2}+1)/(u^{1/2}-1)\rvert$	$\dfrac{\pi^2}{8}\left[\dfrac{\Gamma\left(\dfrac{1}{4}\right)}{\Gamma\left(\dfrac{3}{4}\right)}\right]^2$
(42)	$u(u^2-z^2)^{-1}(1-u^2)^{-1/2}\ln\left(\dfrac{1-z^2}{1-u^2}\right)$	$\dfrac{\pi^2}{8}\mathbf{K}(z)(1-z^2)^{-1/2}$, $\quad 0<z<1$
(43)	$u(z^2-u^2)^{-1}(1-u^2)^{-1/2}\left\{1-\left(\dfrac{1-u^2}{z^2-u^2}\right)^{1/2}\ln\left[\dfrac{(1-u^2)^{1/2}+(z^2-u^2)^{1/2}}{(z^2-1)^{1/2}}\right]\right\}$	$\dfrac{\pi^2}{4z}(z^2-1)^{-1/2}$, $\quad z>1$
(44)	$\int_0^x\left[\dfrac{(u^2+a^2)^{1/2}-a}{u^2+a^2}\right]^{1/2}\mathbf{K}\left[\dfrac{(u^2+1)^{1/2}-1}{(u^2+1)^{1/2}+1}\right]\dfrac{du}{(u^2+1)^{1/2}+1} = [a-(a^2-1)^{1/2}]^{1/2}\operatorname{sech}^2 x\ \mathbf{K}(\operatorname{sech}x)\mathbf{K}(\tanh x)$ $\cosh x = 2^{-1/2}\{1+[2a^2-2a(a^2-1)^{1/2}]^{1/2}\}^{1/2}$, $\operatorname{Re} a\ge 0$ [6]	
(45)	$\int_0^1 u^{2\sigma+1}(2-u^2)^\sigma(1-u^2)^{1/2}\,\mathbf{K}(2^{-1/2}u)du = \dfrac{\pi}{8}\dfrac{1}{\sigma+1}\dfrac{\Gamma\left(\dfrac{3}{4}\right)\Gamma\left(\dfrac{3}{2}+\dfrac{1}{2}\sigma\right)\Gamma\left(1+\dfrac{1}{2}\sigma\right)}{\Gamma\left(\sigma+\dfrac{3}{2}\right)\Gamma^2\left(\dfrac{5}{4}+\dfrac{1}{2}\sigma\right)}$, $\operatorname{Re}\sigma>-1$	

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)du$
(46)	$(1-u)^{-1/2}$	$2\sqrt{2}\,k\,\mathbf{K}(k)\mathbf{K}'(k)$, $k=\sqrt{2}-1$
		SIAM REVIEW **14,** 497 (1972)

(47) $$\int_0^1 \left[u(1-u^2)^{-1/2}\,\mathbf{K}(u)+\ln\left(\frac{1}{4}u\right)\right]\frac{du}{u}=2(\ln 2)^2-\frac{\pi}{12}$$

(48) $$\int_z^\infty u^{-2}(u^{-2}-z^{-2})^{a-1}\,\mathbf{K}(u)\,du=\frac{1}{2}\Gamma(a)e^{\pi i a}z^{-2a}(1-z^2)^{a/2}Q_{-1/2}^{-a}(2z^{-2}-1)$$

$$0<a<\frac{1}{2},\qquad 0<z<1$$

(49) $$\int_\beta^1 \mathbf{K}[(1-\beta)u/\beta][(\beta+1)u-(u^2+\beta)]^{-1/2}\,du=\frac{4}{(1+\beta)\sqrt{z}}\left\{\frac{(z+1)(z+4)\beta}{(z/2)^2+2z+2+(2+z)(z+1)^{1/2}}\right\}^{1/2}\mathbf{K}(k_+)\,\mathbf{K}(k_-)$$

where

$$z=\frac{2[1-\beta+\beta^2]-2[\beta(5\beta-2\beta^2-2)]^{1/2}}{[4\beta-1-\beta^2]+2[\beta(5\beta-2\beta^2-2)]^{1/2}}$$

$$k_\pm=\left\{\frac{z(\sqrt{z+1}\pm\sqrt{z+4})-2(1-\sqrt{z+1})}{z(\sqrt{z+1}\pm\sqrt{z+4})+2(1+\sqrt{z+1})}\right\}^{1/2},\qquad \frac{1}{2}\le\beta\le 1$$

II. Trigonometric integrands

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)du$	
(1)	$u(1-u^2)^{-1/2}\cos[2z(1-u^2)^{1/2}]$	$\frac{\pi^2}{4}J_0^2(z)$	[7]
(2)	$u(1-u^2)^{-1}\cos[2z(1-u^2)^{-1/2}]$	$\frac{\pi^2}{8}[\mathbf{Y}_0^2(z)-\mathbf{J}_0^2(z)]$	
(3)	$u(1-u^2)^{-1}\sin[2z(1-u^2)^{-1/2}]$	$-\frac{\pi^2}{4}J_0(z)Y_0(z)$	[7]
(4)	$(1-u^2)^{-1}\cos[2zu(1-u^2)^{-1/2}]$	$\frac{1}{2}\mathbf{K}_0^2(z)$	[7]
(5)	$u\cos[2z(1-u)^{1/2}]$	$\mathbf{J}_0^2(z)$	[7]
(6)	$u^{-1/2}\left(\frac{1+u}{1-u}\right)\sin[4z\,u^{1/2}(1-u)^{-1}]$	$\frac{\pi}{2}I_0(z)K_0(z)$	[7]
(7)	$u^{-1/2}\left(\frac{1+u}{1-u}\right)\cos[4z\,u^{1/2}(1-u)^{-1}]$	$\frac{1}{2}\mathbf{K}_0^2(z)$	[7]
(8)	$u^{-1}(1-u^2)^{-1/2}\cos[2z\,u^{-1}(1-u^2)^{-1/2}]$	$\frac{\pi}{2}I_0(z)K_0(z)$	[7]
(9)	$(1-u^2)^{-3/2}\sin[2z\,u(1-u^2)^{-1/2}]$	$\frac{\pi}{2}I_0(z)K_0(z)$	[7]
(10)	$(1-u^2)^{-1}(1+u^2)^{-1/2}\cos[2z\,u(1-u^2)^{-1/2}]$	$2^{-3/2}\mathbf{K}_0^2(z)$	[7]
(11)	$(3+u)(1+u^2)^{-3/2}u^{1/2}\cos[2^{3/2}z(1-u)^{-1}(1+u)^{1/2}]$	$2^{1/2}\mathbf{K}_0^2(z)$	[7]
(12)	$(3+u)(1+u^2)^{-3/2}u^{1/2}\sin[2^{3/2}z(1-u)^{-1}(1+u)^{1/2}]$	$2^{-1/2}\pi I_0(z)K_0(z)$	[7]
(13)	$u(1-2u^2)^{\nu-1}\sin[2z(1-2u^2)]\,\theta(2^{-1/2}-u)$	$\frac{2^{\nu-3}\pi^{3/2}z}{\Gamma^2\left(\frac{1}{2}\nu+\frac{1}{4}\right)}{}_2F_3\left(\frac{1}{2}+\frac{1}{2}\nu,\,1+\frac{1}{2}\nu;\,\frac{3}{2},\frac{1}{2}\nu+\frac{5}{4},\frac{1}{2}\nu+\frac{5}{4};\,-z^2\right)$	
(14)	$u(1-2u^2)^{\nu-1}\cos[2z(1-2u^2)]\,\theta(2^{-1/2}-u)$	$\frac{2^{-\nu-3}\pi^{3/2}\Gamma(\nu)}{\Gamma^2\left(\frac{1}{2}\nu+\frac{3}{2}\right)}{}_2F_3\left(\frac{1}{2}\nu,\frac{1}{2}\nu+\frac{1}{2};\,\frac{1}{2},\frac{1}{2}\nu+\frac{3}{4},\frac{1}{2}\nu+\frac{3}{4};\,-z^2\right)$ $\mathrm{Re}\,\nu>0$	
(15)	$(1-u^2)^{-3/2}u\sin\left[z\left(\frac{1+u^2}{1-u^2}\right)\right]$	$\frac{\pi^{3/2}}{8}\frac{\mathbf{S}_{1/2,0}(z)}{z^{1/2}\Gamma^2\left(\frac{3}{4}\right)}$	

III. Exponential and hyperbolic integrands

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)du$
(1)	$e^{-2zu} + u^{-1} e^{-2z/u}$	$\frac{\pi}{2} I_0(z)K_0(z)$
(2)	$u(1-u^2)^{-1} \exp[-2z(1-u^2)^{-1/2}]$	$\frac{1}{2} K_0^2(z)$ [7]
(3)	$(1-u^2)^{-1} \exp[-2zu(1-u^2)^{-1/2}]$	$\frac{\pi^2}{8} [J_0^2(z) + Y_0^2(z)]$
(4)	$u(1-u^2)^{-3/2} \exp[-2zu^2/(1-u^2)]$	$\frac{1}{4} \left(\frac{\pi}{2z}\right)^{1/2} e^z K_0(z)$
(5)	$u(1-u^2)^{-1/2} \exp(zu^2)$	$\frac{\pi^2}{4} e^z \, {}_2F_2 [(1/2), (1/2); 1, 1; -z]$
(6)	$u(1-u^2)^{-3/2} \exp[-4zu^2/(1-u^2)^2]$	$\frac{\pi}{16} z^{-3/4} e^{\frac{1}{2}z} W_{-1/4,0}(z)$
(7)	$u(1-u^2)^{-1} (x^2 - y^2u^2)^{-1/2} \exp\left[-\left(\frac{x^2 - y^2u^2}{1-u^2}\right)\right]$	$\frac{1}{4} (x^2 - y^2)^{-1/2} K_0(x+y)K_0(x-y)$
(8)	$u(1-u^2)^{-3/2} A^{-1} \exp(-A)$ $A = \left[x^2 + y^2 + 2xy \left(\frac{1+u^2}{1-u^2}\right)\right]^{1/2}$	$\frac{1}{4} (xy)^{-1/2} K_0(x)K_0(y)$
(9)	$u^{-2\nu} (1-u^2)^{\nu-1} \exp(-z/2u^2)$	$\frac{1}{4} z^{-\nu} e^{-z} G_{23}^{22} \left(z \middle\vert \begin{matrix} 1 & 1 \\ \nu \; \frac{1}{2} & \frac{1}{2} \end{matrix}\right) \quad 0 < \mathrm{Re}\, \nu < \frac{1}{2}$
(10)	$u^{2\nu-1} (1-u^2)^{-\nu-(1/2)} \exp\left[-\frac{1}{2} z \left(\frac{1+u^2}{1-u^2}\right)\right]$	$\frac{\pi^{1/2} \; z^{-\nu} e^{-z}}{2\,\Gamma(\frac{1}{2} - \nu)} \; G_{23}^{31} \left(z \middle\vert \begin{matrix} 1 & 1 \\ \nu \; \frac{1}{2} & \frac{1}{2} \end{matrix}\right) \quad 0 < \mathrm{Re}\, \nu < \frac{1}{2}$
(11)	$u^{\nu+\frac{1}{2}} (1-2u^2)^{\nu-1} (1-u^2)^{-(2\nu+3)/4} \exp[-z(1-2u^2)/u(1-u^2)^{1/2}] \cdot \theta(2^{-1/2} - u)$	$\cdot \; \frac{2^{\nu-(9/4)}}{\pi^{1/2}} G_{24}^{32} \left(z^2 \middle\vert \begin{matrix} 1 - (\nu/2) \; \frac{1}{2} - (\nu/2) \\ 0 \; \frac{1}{2} \; (1/4) - (\nu/2) \; (1/4) - (\nu/2) \end{matrix}\right)$ $\mathrm{Re}\, \nu > 0$
(12)	$u(1-u^2)^{-1/2} \cosh[2z(1-u^2)^{1/2}]$	$\frac{1}{4} \pi^2 I_0^2(z)$
(13)	$\int_0^\infty \frac{\cos(zu)}{\cosh\left(\left(\frac{1}{2}\right) u\right)} \mathbf{K}\left(\tanh \frac{1}{2} u\right) du = \frac{1}{8\pi} \left\vert \Gamma\left(\frac{1}{4} + \frac{1}{2} iz\right)\right\vert^4$	

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)\,du$
(1)	$\theta(2^{-1/2}-u)u(1-2u^2)^{-5/2}J_0[2zu(1-u^2)^{1/2}/(1-2u^2)]$	$\frac{1}{2}\pi^{1/2}z^{-3/2}e^{-z}$ $\operatorname{Re} z>0$
(2)	$u(1+u^2)(1-u^2)^{-5/2}J_0[2z\,u/(1-u^2)]$	$2^{-3/2}\frac{\pi}{\Gamma^2\left(\frac{1}{4}\right)}z^{-3/2}K_0(z)$
(3)	$u(1-u^2)^{-3/2}J_0[zu/(1-u^2)^{-1/2}]$	$z^{-1}K_0(z)$ $\quad \operatorname{Re} z>0$ [7]
(4)	$u(1-u^2)^{-2}(1+u^2)^{1/2}J_0\left[2z\left(\frac{1+u^2}{1-u^2}\right)\right]$	$-\frac{2^{-1/2}\pi}{16z}[\cos(z)Y_0(z)+\sin(z)J_0(z)]$ $\operatorname{Re} z>0$
(5)	$\theta(2^{-1/2}-u)u(1-2u^2)^{-1/2}J_0[2z(1-2u^2)]$	$\frac{2^{1/2}\pi}{16z}\sin(z)J_0(z)$ $\operatorname{Re} z>0$
(6)	$\int_0^x u^{1/2}\left[\frac{2u}{x^2+(u+1)^2}\right]^{1/2}J_0(xt)\mathbf{K}\left[\left(\frac{2u}{x^2+(u+1)^2}\right)^{1/2}\right]du=\frac{\pi}{x}e^{-xt}J_0(x)$	$x>0$
(7)	$u(1+u^2)^{1/2}(1-u^2)^{-2}Y_0\left[2z\left(\frac{1+u^2}{1-u^2}\right)\right]$	$\frac{2^{-1/2}\pi}{16z}[\cos(z)J_0(z)-\sin(z)Y_0(z)]$ $\operatorname{Re} z>0$
(8)	$u(1-u^2)^{-3/2}K_0[z(1-u^2)^{-1/2}]$	$(\pi/2z)^{3/2}e^{-z}$ $\operatorname{Re} z>0$
(9)	$u(1+u^2)^{1/2}(1-u^2)^{-2}K_0\left[2z\left(\frac{1+u^2}{1-u^2}\right)\right]$	$\frac{2^{-1/2}\pi}{16z}e^{-z}K_0(z)$ $\operatorname{Re} z>0$
(10)	$u(1-u^2)^{-3/2}K_0[zu(1-u^2)^{-1/2}]$	$\frac{\pi^2}{4z}[\mathbf{H}_0(z)-Y_0(z)]$ $\operatorname{Re} z>0$
(11)	$u^2(1-u^2)^{-2}K_0[2z\,u/(1-u^2)]$	$\frac{2^{-1/2}\pi}{8z^{1/2}}S_{-1/2,0}(z)$ $\operatorname{Re} z>0$
(12)	$u^2(1-u^2)^{-5/2}K_0[2zu^2/(1-u^2)]$	$\frac{\pi}{16}z^{-3/2}S_{1/2,0}(z)$ $\operatorname{Re} z>0$
(13)	$u(1-u^2)^{-3/2}\exp[-zu^2/2(1-u^2)]K_0\left[\frac{z}{2(1-u^2)}\right]$	$\frac{\pi}{2}(\pi/z)^{1/2}E_1(z)$ $\operatorname{Re} z>0$
(14)	$u(1-u^2)^{-3/2}S_{\nu,0}[z(1-u^2)^{-1/2}]$	$\frac{\pi}{4z}\left[\frac{\Gamma\left(-\frac{1}{2}\nu\right)}{\Gamma\left(\frac{1}{2}-\frac{1}{2}\nu\right)}\right]^2 S_{\nu+1,0}(z)$ $\operatorname{Re} z>0$
(15)	$u(1+u^2)^{1/2}(1-u^2)^{-2}S_{\nu,1/2}\left[z\left(\frac{1+u^2}{1-u^2}\right)\right]$	$(\pi/2)^{1/2}2^{-\nu}\frac{\Gamma^2\left(-\frac{1}{2}\nu\right)}{16z\,\Gamma\left(\frac{1}{2}-\nu\right)}S_{\nu+1,0}(z)$ $\operatorname{Re} z>0$
(16)	$u(1-u^2)^{-5/4}S_{-1/2,1/2}[2z(1-u^2)^{-1/2}]$	$\frac{2^{-1/2}\pi^3}{16}z^{-1/2}[J_0^2(z)+Y_0^2(z)]$ $\operatorname{Re} z>0$
(17)	$u^{-2}\mathbf{H}_0[zu^{-1}(1-u^2)^{1/2}]$	$-z\,K_0(z)$ $\operatorname{Re} z>0$
(18)	$\theta(u-2^{-1/2})u(2u^2-1)^{-5/2}\mathbf{H}_0[2uz(1-u^2)^{1/2}/(2u^2-1)]$	$-\frac{\pi^{1/2}}{8}z^{-3/2}e^{-z}$ $\operatorname{Re} z>0$

	$f(u)$	$\int_0^1 f(u)\mathbf{K}(u)\,du$
(1)	$u(1-u^2)^{-3/2}\exp\left[z^2\left(\frac{1+u^2}{1-u^2}\right)^2\right]\operatorname{Erfc}\left[z\left(\frac{1+u^2}{1-u^2}\right)\right]$	$\frac{1}{16}\Gamma^2\left(\frac{1}{4}\right)z^{-3/2}e^{(1/2)z^2}W_{1/4,0}(z^2)$ $z \neq 0$
(2)	$u\{\exp[-2z(1-u^2)^{1/2}]Ei[-2z(1-u^2)^{1/2}]$ $+\exp[-2z(1-u^2)^{1/2}]\bar{E}i[2z(1-u^2)^{1/2}]\}$	$-\pi I_0(z)K_0(z)$ $\operatorname{Re} z > 0$
(3)	$u(1-u^2)^{-(1/2)}P_\nu(1-2u^2)$	$\frac{1}{4}\frac{\Gamma^4\left(\nu+\frac{1}{2}\right)}{\Gamma^2(\nu+1)\Gamma(2\nu+1)}$ $\nu > -\frac{1}{2}$
(4)	$\theta(2^{-1/2}-u)u\,P_\nu(1-2u^2)$	$\frac{2^{-1/2}}{4\left(\nu+\frac{1}{2}\right)^2}\left[\frac{\Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)}\frac{\Gamma\left(1+\frac{1}{2}\nu\right)}{\Gamma\left(\frac{1}{2}+\frac{1}{2}\nu\right)}\sin\left(\frac{1}{2}\pi\nu\right)\right.$ $\left.+\frac{\Gamma\left(\frac{3}{4}\right)}{\Gamma\left(\frac{1}{4}\right)}\frac{\Gamma\left(\frac{1}{2}+\frac{1}{2}\nu\right)}{\Gamma\left(1+\frac{1}{2}\nu\right)}\cos\left(\frac{1}{2}\pi\nu\right)\right]$ $\operatorname{Re}\nu > -\frac{1}{2}$
(5)	$u\,P_\nu(1-2u^2)$	$\frac{\pi\cos(\pi\nu)+2\sin(\pi\nu)[\gamma+2\ln 2+\psi(\nu+1)]}{\pi(2\nu+1)^2}$ $\operatorname{Re}\nu > -\frac{1}{2}$
(6)	$u(1-u^2)^{-3/2}Q_\nu\left(\frac{1+u^2}{1-u^2}\right)$	$\frac{\pi}{8}\left(\nu+\frac{1}{2}\right)^2, \quad \operatorname{Re}\nu > -\frac{1}{2}$
(7)	$u^{-2}Q_\nu(2u^{-2}-1)$	$\frac{\psi(\nu+1)-\psi\left(\frac{1}{2}\right)}{(2\nu+1)^2}$ $\operatorname{Re}\nu > -\frac{1}{2}$
(8)	$Q_\nu(2u^2-1)$	$\frac{(\pi/2)\cos(\pi\nu)-[\gamma+2\ln 2+\psi(\nu+1)]}{(2\nu+1)^2}$ $\nu \neq -\frac{1}{2}, -1, -2, \ldots$
(9)	$u^{-1}(1-u^2)^{-1/2}{}_1F_1\left[1;\frac{1}{2};-2z(1-u^2)/u^2\right]$	$\frac{\pi}{4}e^z K_0(z)$ $\operatorname{Re} z > 0$
(10)	$u(1-u^2)^{-1/2}{}_1F_1\left(\frac{3}{2};\frac{1}{2};-2z(1-u^2)\right)$	$\frac{\pi^2}{4}I_0(z)e^{-z}$ $\operatorname{Re} z > 0$
(11)	$u^{-1}(1-u^2)^{-1/2}{}_2F_1\left(\frac{1}{2}\nu+\frac{1}{4},\frac{1}{4}-\frac{1}{2}\nu;\frac{1}{2};1-u^{-2}\right)$	$\frac{\pi}{2}\frac{\Gamma\left(\frac{1}{4}+\frac{1}{2}\nu\right)\Gamma\left(\frac{1}{4}-\frac{1}{2}\nu\right)}{\Gamma\left(\frac{3}{4}+\frac{1}{2}\nu\right)\Gamma\left(\frac{3}{4}-\frac{1}{2}\nu\right)}$
(12)	$\int_0^1(1-u)^{\nu-1}(1-uz)^{-\sigma}\,\mathbf{K}[(uz)^{1/2}]\,{}_2F_1\left[\sigma,-\frac{1}{2};\nu;\frac{(1-u)z}{1-uz}\right]du = \frac{\pi}{2\nu}{}_2F_1\left(\sigma+\frac{1}{2},\frac{1}{2};\nu+1;z\right)$ $\operatorname{Re}\nu > 0$	
(13)	$\int_0^1(1-u)^{\nu-2}K[(uz)^{1/2}]\,{}_2F_1\left[\nu-\frac{1}{2},\nu-\frac{1}{2};\nu-1;x(1-u)\right]du = \frac{\pi}{2(\nu-1)}(1-x)^{1-\nu}\,{}_2F_1\left(\frac{1}{2},\frac{1}{2};\nu;z+x-zx\right)$ $\operatorname{Re}\nu > 1$	

A formula which was obtained as a special case of another in the table is indicated by a reference in parentheses. For example, (II–7) refers to formula (7) of section II dealing with trigonometric integrals. When possible a result related to a published one is indicated by a reference in square brackets.

The author will be grateful for being informed of any errors found in the table, or of additional results.

References

[1] Watson, G. N., Three triple integrals, Q. J. Math. Oxford **10,** 266 (1939).

[2] Morita, T., and Horiguchi, T., Formulas for the lattice Green's functions for the cubic lattices in terms of the complete elliptic integral, J. Phys. Soc. (Japan) **30,** 957 (1971).

[3] Kaplan, E. L., Multiple elliptic integrals, J. Math. and Physics **29,** 69–75 (1950).

[4] Byrd, P. F., and Friedman, M. D., Handbook of Elliptic Integrals for Engineers and Physicists (Springer Verlag, Berlin 1954).

[5] Gradshteyn, I. S., and Ryzhik, I. M., Tables of Integrals Series and Products (Academic Press, 1965).

[6] Glasser, M. L., An elliptic integral identity, Math. Comp. **25,** 533 (1971). That the restriction on a is that given in 1–44, rather than the one in this paper was pointed out by H. Fettis (private communication).

[7] Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G., Tables of Integral Transforms, Vols. 1,2 (McGraw-Hill Book Co., Inc., New York, 1954).

[8] Okui, S., Complete elliptic integrals resulting from infinite integrals of Bessel functions. Nat. Bur. Stand. (U.S.), J. Res. **78B** (Math. Sci. No. 3, 113–134 (July–Sept. 1974).

(Paper 80B2–446)

JOURNAL OF RESEARCH of the National Bureau of Standards — B. Mathematical Sciences
Vol. 80B, No. 2, April-June 1976

Publications of the National Bureau of Standards*

Citations with Selected Abstracts

J. Res. Nat. Bur. Stand. (U.S.), **80A** *(Phys. and Chem.),* No. 2 (Mar.-Apr. 1976), SD Catalog No. C13.22/sec.A:80/2.

Extinction coefficients of NO_2 and N_2O_4, A. M. Bass, A. E. Ledford, Jr., and A. H. Laufer.

A study of the chemiluminescence of the $Pb + O_3$ reactions, M. J. Kurylo, W. Braun, S. Abramowitz, and M. Krauss.

Structure- and solvent-property relationships for the electronic energies of charge-transfer complexes between certain benzene derivatives, H. Argentar.

Bidirectional reflectometry. Part I. A high resolution laser bidirectional reflectometer with results on several optical coatings, J. J. Hsia and J. C. Richmond.

Bidirectional reflectometry. Part II. Bibliography on scattering by reflection from surfaces, J. C. Richmond and J. J. Hsia.

Energy levels, classified lines, and Zeeman effect of neutral thorium, R. Zalubas.

J. Res. Nat. Bur. Stand. (U.S.), **80B** *(Math. Sci.),* No. 1 (Jan.-Mar. 1976), SD Catalog No. C13.22/sec.B:80/1.

Non-negative sums of roots of unity, M. Newman.

Plane-wave scattering-matrix theory of antennas and antenna-antenna interactions: formulation and and application, D. M. Kerns.

Optimal facility location for normally and exponentially distributed points, I. N. Katz and L. Cooper.

Real, 3×3, D-stable matrices, B. E. Cain.

Transient solutions for stratified fluid flows, D. W. Fox.

Equivalence of partitioned matrices, R. B. Feinberg.

Tables of binomial coefficients and Sterling numbers, K. Goldberg, F. T. Leighton, M. Newman, and S. L. Zuckerman.

H116. **American National Standard N540; classification of radioactive self-luminous light sources,** Nat. Bur. Stand. (U.S.). Handb. 116, 15 pages (Jan. 1976) SD Catalog No. C13.11:116, $1.10.

SP428. Parts 1, 2, and 3. **Bibliography of infrared spectroscopy through 1960,** C. N. R. Rao, S. K. Dikshit, S. A. Kudchadker, D. S. Gupta, V. A. Narayan, and J. J. Comeford, Nat. Bur. Stand. (U.S.), Spec. Publ. 428/1-3, 2399 pages (Jan. 1976) SD Catalog No. C13.10:428/1-3, $28.50.

**Publications with prices and SD Catalog numbers may be purchased directly from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (foreign: one-fourth additional). Microfiche copies of the above, plus paper copy and microfiche versions of NBS Interagency Reports (NBSIR's), may be purchased from the National Technical Information Service (NTIS), Springfield, Va. 22161. Prices are subject to change. Reprints from the NBS Journal of Research and non-NBS publications may often be obtained directly from the authors.*

BSS67. **Abnormal loading on buildings and progressive collapse. An annotated bibliography,** E. V. Leyendecker, J. E. Breen, N. F. Somes, and M. Swatta, Nat. Bur. Stand. (U.S.), Bldg. Sci. Ser. 67, 60 pages (Jan. 1976) SD Catalog No. C13.29/2:67, $1.20.

BSS80. **Safety during construction of concrete buildings—A status report,** H. S. Lew, Nat. Bur. Stand. (U.S.), Bldg. Sci. Ser. 80, 56 pages (Jan. 1976) SD Catalog No. C13.29/2:80, $1.15.

TN674. **Report on the 1975 survey of users of the services of Radio Stations WWV and WWVH,** J. A. Barnes and R. E. Beehler, Nat. Bur. Stand. (U.S.), Tech. Note 674, 91 pages (Oct. 1975) SD Catalog No. C13.46:674, $1.50.

TN859. **Literature search: Law enforcement facilities—Planning, design, construction,** R. Kapsch and J. Stroik, Eds., Nat. Bur. Stand. (U.S.), Tech. Note 859, 221 pages (Nov. 1975) SD Catalog No. C13.46:859, $4.15.

TN895. **An overview of floor slip-resistance research with annotated bibliography,** R. J. Brungraber, Nat. Bur. Stand. (U.S.), Tech. Note 895, 113 pages (Jan. 1976) SD Catalog No. C13.46:895, $2.30.

TN896. **NBS Reactor: Summary of activities July 1974 to June 1975,** R. S. Carter, Nat. Bur. Stand. (U.S.), Tech. Note 896, 150 pages (Jan. 1976) SD Catalog No. C13.46:896, $2.15.

NBSIR 73-264. **Examination of failed one inch black iron pipe natural gas service line, Iowa Public Service Company, Eagle Grove, Iowa,** T. R. Shives, 24 pages (Aug. 30, 1973). Order from NTIS as PB243541, $3.50.

NBSIR 74-379. **Completion of the program to evaluate/improve instrumentation and test methods for electroexplosive device safety qualification,** P. A. Hudson, D. G. Melquist, A. R. Ondrejka, and P. E. Werner, 37 pages (June 1974). Order from NTIS as PB247658, $4.00.

NBSIR 74-507. **Mercurous chloride polarizer material,** R. Forman, W. S. Brower, and H. S. Parker, 31 pages (June 1, 1974). Order from NTIS as ADA000343, $4.00.

NBSIR 74-518. **Instrumental colorimetry of retroreflective sign materials,** I. Nimeroff and W. A. Hall, 89 pages (Aug. 1974). Order from NTIS as PB239633, $5.00.

NBSIR 74-580. **Spectral characteristics of additional Bar Code Readers. II,** J. Cohen, 12 pages (Sept. 1974). Order from NTIS as PB248465, $3.50.

NBSIR 74-625. **Examination of failed two inch steel pipe natural gas main, Columbia Gas Company, Spring Garden**

Township, York County, Pennsylvania, T. R. Shives, 41 pages (Jan. 1975). Order from NTIS as PB243547, $4.00.

NBSIR 75-653. **Measurement methodology and supporting documentation for portable air compressor noise,** C. I. Holmer, 48 pages (Jan. 1975). Order from NTIS as PB248097, $4.00.

NBSIR 75-678. **Analysis of construction systems for the thermal classification of residential buildings,** S. T. Liu, 59 pages (Nov. 24, 1975). Order from NTIS as PB248687, $4.50.

NBSIR 75-706. **Calculations of radiant heat flux in the proposed floor covering flame spread test apparatus,** J. Quintiere and K. Bromberg, 21 pages (Dec. 1975). Order from NTIS as PB248640, $3.50.

NBSIR 75-722. **Closure testing equipment instruction manual,** M. S. Morse, 42 pages (Oct. 1975). Order from NTIS as PB248641, $4.00.

NBSIR 75-742. **Mixed oxides for fuel cell electrodes,** U. Bertocci, M. I. Cohen, W. S. Horton, T. Negas, and A. R. Siedle, 56 pages (Jan. 1976). Order from NTIS as PB248744, $4.50.

NBSIR 75-745. **A model for salmon fishery regulatory analysis,** F. C. Johnson, 33 pages (July 10, 1975). Order from NTIS as PB247657, $4.00.

NBSIR 75-755. **The CPSC road test of bicycle braking performance—Experimental evaluation,** D. E. Marlowe, 32 pages (Aug. 1975). Order from NTIS as PB248642, $4.00.

NBSIR 75-757. **A technique for the measurement of flash fire potential of polymeric materials,** J. E. Brown and J. J. Comeford, 33 pages (Dec. 1975). Order from NTIS as PB248914, $4.00.

NBSIR 75-768. **Reevaluation of the densities of the four NBS silicon crystal standards,** H. A. Bowman, R. M. Schoonover, and C. L. Carroll, 36 pages (Aug. 1975). Order from NTIS as PB247943, $4.00.

NBSIR 75-769. **Report on an NBS/AID/OAS Workshop on standardization and measurement services in industrializing economies,** H. S. Peiser, R. S. Marvin, M. McNeil, and J. Mejeur, Eds., 227 pages (Sept. 1975). Order from NTIS as PB247978, $8.00.

NBSIR 75-775. **Electrical measurement of high voltage pulses in diagnostic x-ray units,** R. E. Hebner, Jr., 62 pages (Nov. 1, 1975). Order from NTIS as PB248684, $4.50.

NBSIR 75-782. **Full-scale corridor fire experiment using a glass fiber carpet,** L. H. Breden, 33 pages (Dec. 1975). Order from NTIS as PB248643, $4.00.

NBSIR 75-793. **Microsecond-resolution pulse calorimetry for semiconducting materials at high temperatures (a feasibility study),** A. Cezairliyan and C. W. Beckett, 51 pages (Oct. 1975). Order from NTIS as PB247538, $4.50.

NBSIR 75-796. **An analysis of the Salmon Statistical Information System,** F. C. Johnson, 20 pages (Sept. 29, 1975). Order from NTIS as PB247656, $3.50.

NBSIR 75-818. **Bibliography of the Electromagnetics Division June 30, 1974 to June 30, 1975,** M. L. Woolley, 27 pages (Sept. 1975). Order from NTIS as PB246439, $4.00.

NBSIR 75-820. **Study of cryogenic propellant systems for loading the space shuttle—Part II, Hydrogen systems,** W. G. Steward, 63 pages (Oct. 1975). Order from NTIS as PB246658, $5.00.

NBSIR 75-822. **Study of errors in absolute flux density measurements of Cassiopeia A,** M. Kanda, 34 pages (Oct. 1975). Order from NTIS as PB246933, $4.00.

NBSIR 75-823. **Helium research in support of superconducting power transmission,** V. D. Arp, D. E. Daney, N. V. Frederick, M. C. Jones, P. R. Ludtke, W. R. Parrish, and R. L. Powell, 79 pages (Oct. 1975). Order from NTIS as PB246436, $4.50.

NBSIR 75-910. **The role of chemical kinetics in energy conservation,** H. M. Rosenstock, D. Garvin, J. T. Herron, and W. Tsang, 123 pages (Oct. 1975). Order from NTIS as PB248646, $5.50.

NBSIR 75-917. **A multiple-chamber humidity apparatus,** L. Greenspan, 22 pages (Dec. 1975). Order from NTIS as PB247655, $3.50.

NBSIR 75-920. **Distribution of water through a vertical plane from automatic sprinkler heads,** R. L. P. Custer and K. Wahle, 48 pages (Dec. 1975). Order from NTIS as PB248913, $4.00.

NBSIR 75-924. **Evaluation, revision and application of the NBS stylus/computer system for surface roughness measurement: Minicomputer software,** E. C. Teague, 79 pages (Apr. 1975). Order from NTIS as PB248686, $5.00.

NBSIR 75-952. **An error analysis of failure prediction techniques derived from fracture mechanics,** S. M. Wiederhorn, E. R. Fuller, Jr., J. Mandel, and A. G. Evans, 41 pages (Dec. 1975). Order from NTIS as PB248910, $4.00.

NBSIR 75-955. **Durability and related tests for selected elements and materials used in the exterior envelope of buildings,** L. W. Masters, E. J. Clark, G. A. Sleater, and A. Hockman, 151 pages (Nov. 1975). Order from NTIS as PB248685, $6.75.

NBSIR 75-956. **An interference simulator for quantitative determination of the susceptibility of flame detectors to false alarm,** A. Scheidweiler, 16 pages (Nov. 1975). Order from NTIS as PB247654, $3.50.

NBSIR 75-958. **Simulated precipitation reference materials,** J. K. Taylor, E. R. Deardorff, R. A. Durst, E. J. Maienthal, T. C. Rains, and E. P. Scheide, 25 pages (Oct. 1975). Order from NTIS as PB248743, $3.50.

NBSIR 75-962. **An outdoor noise monitoring system with automatic calibration and remote digital display,** D. S. Blomquist, J. S. Forrer, and D. M. Corley, 11 pages (Oct. 1975). Order from NTIS as PB247639, $3.50.

NBSIR 76-982. **Data processing and data analysis procedures for fire load and live load survey program,** R. M. McCabe, C. Culver, L. T. Lee, and J. G. Hirschberg, 344 pages (Dec. 1975). Order from NTIS as PB248699, $10.00.

This column lists all outside publications by the NBS staff, as soon as possible after issuance as practical. For completeness, earlier references not previously reported may be included from time to time.

Abrams, M. D., **Computer communications network performance measurement (Abstract),** *Proc. Eurocon 1974 Conf. Digest on Electrotechnics, Amsterdam, The Netherlands, Apr. 22-26, 1974,* 2 pages (1974).

Abrams, M. D., **Consumer-oriented measurement of computer network performance,** *Proc. IEEE National Telecommunications Conf., San Diego, Calif., Dec. 2-4, 1974,* pp. 843-844 (1974).

Achenbach, P. R., Didion, D. A., **Energy conservation measures in the National Bureau of Standards laboratory complex,** *Proc. Energy Conservation and Energy Management in Buildings, London, England, Nov. 13-14, 1975,* pp. 51-68 (1975).

Achenbach, P. R., **National Bureau of Standards research in support of energy conservation standards for buildings,** *Proc. 10th Intersociety Energy Conversion Engineering Conf., Newark, Del., Aug. 18-22, 1975*, pp. 667-676 (1975).

Allan, D. W., Daams, H., **Picosecond time difference measurement system,** *Proc. 29th Annual Symp. on Frequency Control, Atlantic City, N.J., May 28-30, 1975*, pp. 404-411 (1975).

Allan, D. W., Hellwig, H., Glaze, D. J., **An accuracy algorithm for an atomic time scale,** *Metrologia* **11**, No. 3, 133-138 (1975).

Andrews, J. R., Baldwin, E. E., **Baseband impulse generator useful to 5 GHz,** *Proc. 1975 IEEE Int. Symp. on Electromagnetic Compatibility, San Antonio, Tex., Oct. 7-9, 1975*, pp. 1-4 (1975).

Andrews, J. R., **Directional-coupler technique for triggering a tunnel diode,** *IEEE Trans. Instrum. Meas.* **IM-24**, No. 3, 275-277 (Sept. 1975).

Andrews, J. R., Gans, W. L., **Pulsed wavemeter timing reference for sampling oscilloscope calibration,** *IEEE Trans. Instrum. Meas. Short Papers*, **IM-24**, No. 1, 82 (Mar. 1975).

Andrews, J. R., Gans, W. L., **Time domain automatic network analyzer,** (Proc. Colloque International sur L'Electronique et la mesure, Paris, France, May 26-30, 1975), Paper in *La Mesure Electrique de Precision*, pp. 258-267 (Le Comité d'Organisation du Colloque International sur L'Electronique et la mesure, Paris, France, 1975).

Ausloos, P., **Pulse radiolysis of alkanes in the gas-phase, ion-molecule reactions and neutralization mechanism of hydrocarbon ions,** (Proc. Symp. on Mechanisms of Hydrocarbon Reactions, Siófok, Hungary, June 5-7, 1973), Paper in *Mechanisms of Hydrocarbon Reactions*, F. Márta and D. Kalló, Eds., *Plenary Lecture* **3**, pp. 603-624 (1975).

Beaty, E. C., **Measurements of the energy and angular distribution of secondary electrons,** *Radiat. Res.* **64**, 70-79 (1975).

Beers, Y., Howard, C. J., **The microwave spectrum of HO_2 near 65 GHz,** *J. Chem. Phys.* **63**, No. 10, 4212-4216 (Nov. 15, 1975).

Bender, P. L., **Laser measurements of the lunar distance,** *Rev. Geophys. Space Phys.* **13**, No. 3, 271-272, 291 (July 1975).

Berg, J. L., **Federal Privacy Act countdown—Is the private sector ready?,** *Infosystems* **22**, No. 7, 25-27 (July 1975).

Berg, N. J., Lieberman, A. G., **The effects of radiation-induced displacement damage on impurity conduction in gallium arsenide,** *J. Appl. Phys.* **46**, No. 8, 3475-3482 (Aug. 1975).

Berger, M. J., Seltzer, S. M., Domen, S. R., Lamperti, P. J., **Stopping-power ratios for electron dosimetry with ionization chambers,** (Proc. Int. Symp. on Advances in Biomedical Dosimetry, Vienna, Austria, Mar. 1975), Paper in *Advances in Biomedical Dosimetry*, pp. 589-609 (International Atomic Energy Agency, Vienna, Austria, 1975).

Berke, J. G., **Procurement—The pull side of technology transfer,** (Proc. ASME 5th Intersociety Conf. on Environmental Systems, San Francisco, Calif., July 21-24, 1975), *75-ENAs-3*, pp. 1-4 (The American Society of Mechanical Engineers, New York, N.Y., 1975).

Birky, M. M., **Review of smoke and toxic gas hazards in fire environment,** *Proc. Int. Symp. on Fire Safety of Combustible Materials, Edinburgh, Scotland, Oct. 15-17, 1975*, **1**, 231-252 (1975).

Birmingham, B. W., Smith, C. N., **Cryogenics and the energy crisis,** *Cryogenics* **15**, No. 3, 115-118 (Mar. 1975).

Blanc, R. P., **Availability and useability of computer communication networks,** *Proc. Seventh Int. Conf. on Systems Sciences, Honolulu, Hawaii, Jan. 8-10, 1974*, pp. 10-13 (1974).

Branstad, D. K., **Encryption protection in computer data communications,** *Proc. Fourth Data Communications Symp., Quebec City, Canada, Oct. 7-9, 1975*, pp. 8-1—8-7 (1975).

Buchbinder, B., Buchbinder, L. B., **The fire hazard analysis process at the National Bureau of Standards,** *Proc. Int. Symp. on Fire Safety of Combustible Materials, Edinburgh, Scotland, Oct. 15-17, 1975*, **1**, 156-161 (1975).

Buchbinder, B., Mathers, W., **Preliminary indications from survey of U.S. household fire experience,** *Proc. 8th Annual Meeting Information Council on Fabric Flammability, New York, N.Y., Dec. 5, 1974*, pp. 174-178 (1975).

Buchbinder, B., **Pilot implementation of the Fire Incident System,** *Fire J.* **69**, No. 3, 65-69 (May 1975).

Bussey, H. E., Richmond, J. H., **Scattering by a lossy dielectric circular cylindrical multilayer, numerical values,** *IEEE Trans. Antennas Propag.* **23**, No. 5, 723-725 (Sept. 1975).

Cabana, A., Laurin, M., Lafferty, W. J., Sams, R. L., **High resolution infrared spectra of the ν_2 and $2\nu_1$ bands of $^{14}N^{16}O_2$,** *Can. J. Phys.* **53**, No. 19, 1902-1926 (1975).

Cali, J. P., **Reference materials in clinical chemistry,** *Fed. Proc.* **34**, No. 12, 2123-2126 (Nov. 1975).

Casella, R. C., Rowe, J. M., Trevino, S. F., **Determining the number of independent real parameters in the phonon dynamical matrix,** *Phys. Rev. B* **12**, No. 10, 4573-4574 (Nov. 15, 1975).

Cataland, G., Plumb, H. H., **Fixed points: Superconductive transition temperatures of lead and indium,** *Metrologia* **11**, 161-163 (1975).

Cezairliyan, A., Coslovi, L., Righini, F., Rosso, A., **Radiance temperature of molybdenum at its melting point,** (Proc. Symp. European Conf. on Temperature Measurement, Teddington, England, Apr. 9-11, 1975), Chapter 4 in *Temperature Measurement*, B. F. Billing and T. J. Quinn, Eds., Inst. Phys. Conf. Ser. No. 26, pp. 287-296 (Institute of Physics, London, England, 1975).

Cezairliyan, A., McClure, J. L., **A subsecond pulse heating technique for the study of solid-solid phase transformations at high temperatures: Application to iron,** *High Temp. Sci.* **7**, 189-196 (1975).

Cezairliyan, A., Righini, F., **Measurement of melting point, radiance temperature (at melting point), and electrical resistivity (above 2,100 K) of zirconium by a pulse heating method,** *Rev. Int. Hautes Temp. Refract.* **12**, No. 3, 201-207 (1975).

Chow, L. C., Brown, W. E., **Topical fluoridation of teeth before sealant application,** *J. Dent. Res.* **54**, No. 5, 1089 (Sept.-Oct. 1975).

Chwirut, D. J., **Tensile creep of angle-plied boron/epoxy laminates,** *J. Test. Eval.* **3**, No. 6, 442-448 (1975).

Clifton, J. R., Foster, B. E., Trattner, E., Clevenger, R. A., **Dimensional stability of masonry walls,** *Am. Soc. Test Mater. Spec. Tech. Publ.* **589**, pp. 42-75 (1975).

Clough, R. B., **The effects of specimen size and microstructure on the Larson-Miller parameter,** *Scr. Metall.* **9,** No. 12, 1325-1329 (1975).

Cotton, I. W., **Cost-benefit analysis of interactive systems,** *Proc. Second Jerusalem Conf. on Information Technology, Jerusalem, Israel, July 29-Aug. 1, 1974,* pp. 729-746 (1974).

Cotton, I. W., **Remark on stably updating mean and standard deviation of data,** *Commun. ACM* **18,** No. 8, 458 (Aug. 1975).

Crawford, E. A., Phelps, A. V., **Formative time lags in CO_2 laser discharges,** *Appl. Phys. Lett.* **25,** No. 1, 59-61 (July 1, 1974).

Cullen, W. C., **The composite roofing membrane,** *The Roofing Spec.* **3,** No. 5, 17-21 (Sept. 1975).

Danos, M., **Relativistic nuclear physics, why and how?,** (Proc. Int. Conf. on Interaction Studies in Nuclei, Mainz, Germany, Feb. 17-20, 1975), Paper in *Interaction Studies in Nuclei,* H. Jochim and B. Ziegler, Eds., pp. 885-910 (North-Holland Publishing Co., The Netherlands, 1975).

Daywitt, W. C., Kanda, M., **G/T measurement errors with radio stars,** *Proc. IEEE Int. AP-Symp., Urbana/Champaign, Ill., June 2-5, 1975, Session* **20,** pp. 460-463 (1975).

Drullinger, R. E., Hessel, M. M., Smith, E. W., **New laser measurement techniques for excited electronic states of diatomic molecules,** *Proc. Megeve Laser Spectroscopy Conf., Megeve, France, June 23-27, 1975, Lecture Notes in Physics,* 91-99 (1975).

Dunn, G. H., **Collision studies with ion storage techniques,** (Proc. Int. Conf. on Atomic Physics, Heidelberg, Germany, July 1974), Paper in *Atomic Physics* **4,** pp. 575-587 (Plenum Press, New York, N.Y., 1975).

Dziuba, R. F., Field, B. F., Finnegan, T. F., **Cryogenic voltage comparator system for 2e/h measurements,** *IEEE Trans. Instrum. Meas.* **IM-23,** No. 4, 264-267 (Dec. 1974).

Dziuba, R. F., Sullivan, D. B., **Cryogenic direct current comparators and their applications,** *IEEE Trans. Magn.* **MAG-11,** No. 2, 716-719 (Mar. 1975).

Engen, G. F., **Automated calibration of directional-coupler-bolometer-mount assemblies,** *IEEE Trans. Microwave Theory Tech.* **MTT-23,** No. 12, 984-990 (Dec. 1975).

Engen, G. F., **Automated calibration of directional-coupler-bolometer-mount assemblies,** *Proc. IEEE-MTT-S Int. Microwave Symp., Palo Alto, Calif., May 12-14, 1975,* pp. 85-97 (Institute of Electrical and Electronics Engineers, Inc., Piscataway, N.J., May 1975).

Evenson, K. M., **Frequency measurements in the optical region and the speed of light,** (Proc. ICO Conf. Optical Methods in Science and Industrial Measurements, Tokyo, Japan, Aug. 26-30, 1974), *Japan J. Appl. Phys.* **14,** Suppl. 14-1, 1-10 (1975).

Evenson, K. M., Howard, C. J., **Laser magnetic resonance spectroscopy,** *Laser Spectrosc.,* pp. 535-540 (1975).

Evenson, K. M., Petersen, F. R., **Stabilized lasers and applications,** *Proc. Physics of Quantum Electronics Summer School, Crystal Mountain, Wash., July 8-20, 1973,* pp. 367-436 (1975).

Evenson, K. M., Petersen, F. R., Wells, J. S., **Speed of light from direct laser frequency and wavelength measurements: Emergence of a laser standard of length,** *Laser Spectrosc.,* pp. 143-146 (1975).

Fechter, J. V., Pezoldt, V. J., Persensky, J. J., Lepkowski, J. R., **Study of the National Standards for Directional and Other Official Signs. Overview of their adequacy,** *Report No. FHWA-RD-75-118,* 75 pages (Available from the National Technical Information Service, Springfield, Va., 22161, Oct. 1975).

Finnegan, T. F., Toots, J., Wilson, J., **Frequency-pulling and coherent-locking in thin-film Josephson oscillators,** (Proc. 14th Int. Conf. on Low Temperature Physics, Helsinki, Finland, Aug. 14-20, 1975), Paper in *Low Temperature Physics* **LT-14,** N. Krusius and N. Vuorio, Eds., **4,** 184-187 (North Holland Publishing Co., Amsterdam, 1975).

Fong, E., **A benchmark test approach for generalized data base software,** *Proc. COMPCON fall Conf. Digest, Washington, D.C., Sept. 8-10, 1975,* 7 pages (1975).

Forman, R. A., Thurber, W. R., Aspnes, D. E., **Second indirect band gap in silicon,** *Solid State Commun.* **14,** No. 10, 1007-1010 (1974).

Gallagher, A., **Noble-gas broadening of the Li resonance line,** *Phys. Rev. A* **12,** No. 1, 133-138 (July 1975).

Gebbie, K. B., Steinitz, R., **Comparison of Hα and Ca II H and K spectroheliograms as a diagnostic probe,** (Proc. Int. Astronomical Union Symp. on Chromospheric Fine Structure, Surfers' Paradise, Qld. Australia, Sept. 1973), Paper in *Chromospheric Fine Structure,* R. Grant Athay, Ed., No. 56, 55-63 (Reidel Publishing Co., Boston, Mass., 1974).

Glaze, D. J., Hellwig, H., Allan, D. W., Jarvis, S., Jr., Wainwright, A. E., **Accuracy evaluation and stability of the NBS primary frequency standards,** *IEEE Trans. Instrum. Meas.* **IM-23,** 489-501 (Dec. 1974).

Goldman, A. J., **The adequacy of management science technology for nonmilitary applications in the Federal Government,** (Proc. Workshop on Management Science in the Federal Government, Washington, D.C., Sept. 18, 1969), Chapter 6 in *Management and Policy Science in American Government,* M. J. White, M. Radnor, and D. A. Tansik, Eds., Part III, pp. 135-170 (Lexington Books, D. C. Heath & Co., Lexington, Mass., 1975).

Graminski, E. L., **Problems and potentials in paper recycling,** (Proc. National Materials Conservation Symp. on Resource Recovery and Utilization, Gaithersburg, Md., Apr. 29-May 1, 1974), Paper in *Resource Recovery and Utilization,* H. Alter and E. Horowitz, Eds., *Am. Soc. Test. Mater. Spec. Tech. Publ.* **592,** pp. 132-139 (American Society for Testing and Materials, Philadelphia, Pa., 1975).

Gravatt, C. C., Allegrini, I., **A new light scattering method for the determination of the size distribution of particulate matter in air,** (Proc. 3rd Int. Clean Air Congress (IUAPPA), Dusseldorf, W. Germany, Oct. 8-12, 1973), Paper in *Proceedings of the 3rd International Clean Air Congress,* pp. C3-C5 (Verlag GMBH, Dusseldorf, W. Germany, 1973).

Greenberg, O. W., Nelson, C. A., **Composite models of leptons,** *Phys. Rev. D* **10,** No. 8, 2567-2573 (Oct. 15, 1974).

Greer, S. C., Hocken, R., **Thermal expansion near a critical solution point,** *J. Chem. Phys.* **63,** No. 12, pp. 5067-5072 (Dec. 15, 1975).

Hamilton, C. A., Phelan, R. J., Jr., Day, G. W., **Pyroelectric radiometers,** *Opt. Spectra* **9,** No. 10, 37-38 (Oct. 1975).

Hammond, H. K., Hsia, J. J., **Evaluation of instrument tolerances for 75° gloss,** *Tappi Notes to Editor* **58,** No. 11, 143-144 (Nov. 1975).

Hellwig, H., Wainwright, A. E., **Submicrosecond time transport with a rubidium portable clock,** *Proc. 29th Annual Symp. on Frequency Control, Atlantic City, N.J., May 28-30, 1975,* pp. 384-386 (1975).

Hessel, M. M., Drullinger, R. E., Broida, H. P., **Chemiluminescent reactions in a heat-pipe oven,** *J. Appl. Phys.* **46,** No. 5, 2317-2318 (May 1975).

Hiza, M. J., Kidnay, A. J., Miller, R. C., **Equilibrium properties of fluid mixtures, a bibliography of data on fluids of cryogenic interest,** Volume in *NSRDS Bibliographic Series,* 160 pages (IFI/Plenum Data Co., New York, N.Y., 1975).

Hocken, R., Moldover, M. R., Muth, E., Gerner, S., **Versatile cells for optical studies in fluids,** *Rev. Sci. Instrum.* **46,** No. 12, 1699-1700 (Dec. 1975).

Hoer, C. A., Roe, K. C., **Using an arbitrary six-port junction to measure complex voltage ratios,** *IEEE Trans. Microwave Theory Tech.* **MTT-23,** No. 12, 978-984 (Dec. 1975).

Hoer, C. A., Roe, K. C., **Using an arbitrary six-port junction to measure complex voltage ratios,** *Proc. 1975 IEEE Microwave Theory and Technique Symp. Int., Palo Alto, Calif., May 12-14, 1975,* pp. 98-99 (1975).

Holt, D. R., Nahman, N. S., **Coaxial line pulse-response error due to a planar skin effect approximation,** *Proc. 1972 CPEM Conf. on Precision Electromagnetic Measurement, Boulder, Colo., June 26-29, 1972,* pp. 1-2 (1972).

Holton, J. K., Driscoll, P., **Linking research and practice,** *AIA J.,* pp. 67-68 (Apr. 1975).

Howe, D. A , Salazar, H. F., **A digital 5.00688 MHz synthesizer and squarewave FM servo system for cesium standards,** *Proc. 29th Annual Symp. on Frequency Control, Atlantic City, N.J., May 28-30, 1975,* pp. 387-393 (1975).

Hudson, R. P., Pfeiffer, E. R., **Magnetic susceptibility of single-crystal CMN, 0.002K to 2K,** (Proc. 14th Int. Conf. on Low Temperature Physics, Otaniemi, Finland, Aug. 14-20, 1975), Paper in *Low Temperature Physics* **LT-14,** M. Krusius and M. Vuorio, Eds., **3,** 242-245 (North Holland Publishing Co., Amsterdam, 1975).

Hughes, C. E., Singletary, W. E., **Triadic partial implicational propositional calculi,** *Zeitschr. f. math. Logik und Grundlagen d. Math.* **21,** 21-28 (1975).

Hughes, E. E., Taylor, J. K., **Accurate gas standards for air pollution analyses,** *Proc. WMO/WHO Tech. Conf. on Observation and Measurement of Atmospheric Pollution, Helsinki, Finland, July 30-Aug. 4, 1973,* pp. 1-13 (World Meteorological Organization, Geneva, Switzerland, 1974).

Interrante, C. G., **Report on the ferrous metals workshop,** (Proc. National Materials Conservation Symp. on Resource Recovery and Utilization, Gaithersburg, Md., Apr. 29-May 1, 1974), Paper in *Resource Recovery and Utilization,* H. Alter and E. Horowitz, Eds., *Am. Soc. Test. Mater. Spec. Tech. Publ.* **592,** pp. 146-152 (American Society for Testing and Materials, Philadelphia, Pa., 1975).

Jackson, R. H. F., **Towards algorithmic standardization in mathematical programming: Another step,** *Proc. 1974 Fall UNIVAC Users' Conf., Toronto, Canada, Sept. 1974,* pp. 3-147—3-153 (1974).

Kamas, G., **TV color frequencies defended as accurate,** *Electron. Design Letter to the Editor* **13,** 7 (June 21, 1975).

Kan, P. T., Peterson, G. A., Webb, D. V., Szalata, Z. M., O'Connell, J. S., Fivozinsky, S. P., Lightbody, J. W., Jr., Penner, S., **Electrodisintegration of ^{3}He,** *Phys. Rev. C* **12,** No. 4, 1118-1125 (Oct. 1975).

Kanda, M., **Accuracy considerations in the measurement of the power gain of a large microwave antenna,** *IEEE Trans. Antennas Propag.* **AP-23,** No. 3, 407-411 (May 1975).

Kanda, M., **A measure for the stability of solid state noise sources,** *Proc. 1975 IEEE Microwave Theory and Technique Symp. Int., Palo Alto, Calif., May 12-14, 1975,* pp. 315-317 (1975).

Kasen, M. B., **Mechanical and thermal properties of filamentary-reinforced structural composites at cryogenic temperatures. 2: Advanced composites,** *Cryogenics* **15,** No. 12, 701-722 (Dec. 1975).

King, G. C., Read, F. H., Bradford, R. C., **Structure near autoionizing energies in the excitation of bound states of helium, neon, and argon by electron impact,** *J. Phys. B* **8,** 2210 (1975).

Koyanagi, R. S., **Development of a low-frequency-vibration calibration system,** *Exp. Mech.* **15,** 443-448 (Nov. 1975).

Krauss, M., Neumann, D., **Ion-pair states of O_2,** *J. Chem. Phys.* **63,** No. 12, 5073-5076 (Dec. 15, 1975).

Krauss, M., Neumann, D., **On the interaction of $O(^1S)$ with $O(^3P)$,** *Chem. Phys. Lett.* **36,** No. 3, 372-374 (Nov. 15, 1975).

Krauss, M., Neumann, D., **The dipole moment function of $CO(a^3\Pi)$,** *Mol. Phys.* **30,** No. 4, 1015-1020 (1975).

Kunasz, P. B., Hummer, D. G., Mihalas, D., **Theory of extended stellar atmospheres. II. A grid of static spherical models for O stars and planetary nebula nuclei,** *Astrophys. J.* **202,** 92-113 (Nov. 1975).

Lawton, R. A., Andrews, J. R., **Pulsed-laser application to sampling oscilloscope,** *Electron. Lett.* **11,** No. 7, 138 (Apr. 3, 1975).

Lawton, R. A., Scavannec, A., **Photoconductive detector of fast-transition optical waveforms,** *Electron. Lett.* **11,** No. 4, 74-75 (Feb. 20, 1975).

Lee, T. G., Parker, W. J., Tryon, M., **Laboratory fire performance characteristics of a dibromotetrafluoroethane-blown rigid polyurethane foam,** *J. Fire Flammability* **6,** 499-510 (Oct. 1975).

Lovas, F. J., **Application of microwave spectroscopy to chemical analysis,** *ISA Trans.* **14,** No. 2, 145-151 (1975).

Lyon, G., Walker, J. C., **On some polynomial search methods for hash tables of prime and composite sizes,** *Proc. Canadian Information Processing Society (CIPS) Nat. Conf., Regina, Sask., Canada, June 25, 1975,* pp. 290-299 (1975).

Masters, L. W., Reichard, T. W., **Evaluation of adhesive-bonded joints in housing components of glass fiber-reinforced polyester laminate,** (Proc. American Society for Testing and Materials Conf. on Composite Reliability, Las Vegas, Nev., Apr. 15-16, 1974), Paper in *Composite Reliability, Am. Soc. Test. Mater. Spec. Tech. Publ.* **580,** pp. 146-156 (American Society for Testing and Materials, Philadelphia, Pa., Aug. 1975).

Mathey, R. G., Cullen, W. C., **Performance criteria for built-up roofing—Part 2,** *Roofing Siding Insulation* **52,** No. 1, 44-46, 56 (Jan. 1975).

May, W. E., Chesler, S. N., Cram, S. P., Gump, B. H., Hertz, H. S., Enagonio, D. P., Dyszel, S. M., **Chromatographic analysis of hydrocarbons in marine sediments and seawater,** *J. Chromatogr. Sci.* **13,** 535-540 (Nov. 1975).

McCaffrey, B. J., Quintiere, J. G., **Fire-induced corridor flow in a scale model study,** *Proc. International Council on Building Research and Documentation (CIB) Symp. on the Control of Smoke Movement in Building Fires, Garston, Watford, England, Nov. 4-5, 1975,* **1,** 34-47 (1975).

McCamy, C. S., **Specification of geometric and spectral conditions relating to densitometry,** Paper 15.2 in *SPSE Handbook of Photographic Science and Engineering,* W. Thomas, Jr., Ed., pp. 831-840 (John Wiley & Sons, New York, N.Y., 1973).

McClintock, W., Linsky, J. L., Henry, R. C., Moos, H. W., Gerola, H., **Ultraviolet observations of cool stars. III. Chromospheric and coronal lines in α Tauri, β Geminorum, and α Bootis,** *Astrophys. J.* **202,** 165-182 (Nov. 1975).

Mighell, A. D., Santoro, A., **Geometrical ambiguities in the indexing of powder patterns,** *J. Appl. Crystallogr.* **8,** Part 3, 372-374 (June 1975).

Mighell, A., Santoro, A., Prince, E., Reimann, C., **Neutron diffraction structure determination of dichlorotetrapyrazolecopper(II), $Cu(C_3H_4N_2)_4Cl_2$,** *Acta Crystallogr.* **B31,** Part 10, 2479-2482 (Oct. 15, 1975).

Miller, A., Hussmann, E. K., McLaughlin, W. L., **Interferometer for measuring fast changes of refractive index and temperature in transparent liquids,** *Rev. Sci. Instrum.* **46,** No. 12, 1635-1638 (Dec. 1975).

Mount, G. H., Linsky, J. L., **A new solar carbon abundance based on non-LTE CN molecular spectra,** *Astrophys. J.* **202,** No. 1, L51-L54 (Nov. 15, 1975).

Newell, A. C., **Improved polarization measurements using a modified three antenna technique,** *Proc. IEEE Int. AP-Symp., Urbana/Champaign, Ill., June 2-4, 1975, Session* **15,** pp. 337-340 (1975).

Newell, A. C., Yaghjian, A. D., **Study of errors in planar near-field measurements,** *Proc. IEEE Int. AP-Symp., Urbana/Champaign, Ill., June 2-4, 1975, Session* **20,** pp. 470-473 (1975).

Newman, M., Sheingorn, M., **Continuous solutions of a homogeneous functional equation,** *Aequationes Math.* **13,** No. 1/2, 47-59 (1975).

Key words: analytic functions; functional equations; Hurwitz's theorem.

It is shown that the functional equation

$$f(x)+f(ax)+f(bx)=0 \quad , \quad 1<a<b \quad ,$$

has a nontrivial solution which is continuous for all real x if and only if $b \neq a^2$. Generalizations for the equation

$$\Sigma f(a_k x)=0 \quad , \quad 1=a_1<a_2<\ldots<a_{n_{\max}},$$

are also derived.

Ogburn, F., **Coating thickness—its measurement and its significance,** Chapter 14 in *Properties of Electrodeposits—Their Measurement and Significance,* pp. 229-245 (Electrochemical Society, Princeton, N.J., 1975). 340

Ogburn, F., **Density of electrodeposited metal, its significance and measurement,** Chapter 5 in *Properties of Electrodeposits Their Measurement and Significance,* R. Sard, H. Leidheiser, Jr., and F. Ogburn, Eds., pp. 71-79 (The Electrochemical Society, Inc., Princeton, N.J., 1975).

Ogburn, F., Johnson, C. E., **Mechanical properties of electrodeposited brass,** *Plating* **62,** 141-148 (Feb. 1975).

Ott, W. R., Slater, J., Cooper, J., Gieres, G., **H^- shape-resonance studies in an arc plasma,** *Phys. Rev. A* **12,** No. 5, 2009-2016 (Nov. 1975).

Parrish, W. R., Steward, W. G., **Vapor-liquid equilibria data for helium-carbon monoxide and helium-nitrous oxide systems,** *J. Chem. Eng. Data* **20,** No. 4, 412-416 (Oct. 1975).

Payne, B. F., Koyanagi, R. S., Federman, C., Jones, E., **Accelerometer calibration at the National Bureau of Standards,** *21st Int. Instrumentation Symp. ASD/TMD, Philadelphia, Pa., May 19-21, 1975,* pp. 1-17 (1975).

Pella, P. A., Hughes, E. E., Taylor, J. K., **Development of gas-blending systems for calibration: Application to hydrogen fluoride, arsine and phosgene in air,** *Am. Ind. Hyg. Assoc. J.* **36,** No. 10, 755-759 (Oct. 1975).

Peterlin, A., **Plastic deformation of polymers with fibrous structure,** *Colloid Polym. Sci.* **253,** No. 10, 809-823 (1975).

Peterlin, A., **Structural model of mechanical properties and failure of crystalline polymer solids with fibrous structure,** *Int. J. Fract.* **11,** No. 5, 761-780 (Oct. 1975).

Pontius, P. E., **Mass measurement: A study of anomalies,** *Science* **190,** 379-380 (Oct. 24, 1975).

Pyke, T. N., Jr., **Networked minis and micros—Configurations, applications, and standards,** *Proc. Mini Micro Computer Symp., U.S. Naval Academy, Annapolis, Md., Apr. 15, 1975,* pp. 44-46 (1975).

Quintiere, J., **The application and interpretation of a test method to determine the hazard of floor covering fire spread in building corridors,** *Proc. Int. Symp. on Fire Safety of Combustible Materials, Edinburgh, Scotland, Oct. 15-17, 1975,* **1,** 355-367 (1975).

Rains, T. C., **Iron, cobalt, and nickel,** Chapter 10 in *Flame Emission and Atomic Absorption Spectrometry,* J. A. Dean and T. C. Rains, Eds., **3,** 216-246 (Marcel Dekker, Inc., New York, N.Y., Nov. 1975).

Read, F. H., **Displaced electron energies and the "shake-down" effect,** *Radiat. Res.* **64,** 23-36 (1975).

Rebbert, R. E., Ausloos, P. J., **Photodecomposition of $CFCl_3$ and CF_2Cl_2,** *J. Photochem.* **4,** 419-434 (1975).

Rensberger, R. A., **Lead paint hazard abatement,** *Constr. Specifier* **28,** No. 6, 45-46, 48-49 (June 1975).

Roberts, R. W., **The other face of the measurement base,** *Anal. Chem.* **47,** No. 7, 648A-656A (June 1975).

Rosasco, G. J., Prask, H. J., **Polarized Raman study of lattice modes in ammonium perchlorate at low temperatures,** *Solid State Commun.* **16,** No. 1, 135-138 (1975).

Rosasco, G. J., Simmons, J. H., **Identification of sulfur deposits in bubbles in glass by Raman scattering spectroscopy,** *Am. Ceramic Soc. Bull.* **54,** No. 6, 590-591 (June 1975).

Rosasco, G. J., Simmons, J. H., **Investigation of gas content of inclusions in glass by Raman scattering spectroscopy,** *Am. Ceramic Soc. Bull.* **53,** No. 9, 626-630 (Sept. 1974).

Rothschild, W. G., Rosasco, G. J., Livingston, R. C., **Dynamics of molecular reorientational motion and vibrational relaxation in liquids. Chloroform,** *J. Chem. Phys.* **62,** No. 4, 1253-1268 (Feb. 15, 1975).

Rubin, R. J., Mazur, J., **Ordered spans of unrestricted and self-avoiding random-walk models of polymer chains. I. Space-fixed axes,** *J. Chem. Phys.* **63,** No. 12, 5362-5374 (Dec. 15, 1975).

Rubin, S., **Thermal resistance measurements on monolithic and hybrid Darlington power transistors,** (Proc. 1975 IEEE Power Electronics Specialists Conf., Culver City, Calif., June 9-11, 1975), *PESC 75 Record,* **75-CHO,** 965-4 AES, 252-261 (The Institute of Electrical and Electronics Engineers, New York, N.Y., Sept. 1975).

Saltman, R. G., **The human side of automating,** *Proc. Association for Computing Machinery and National Bureau of Standards Conf., on The Systems Approach: Key to Successful Computer Applications, Gaithersburg, Md., June 20, 1974,* pp. E.4.1-4.11 (1974); *Cyberdent* **2,** No. 3, 1, 3-4 (1974).

Sanchez, I. C., Eby, R. K., **Thermodynamics and crystallization of random copolymers,** *Macromolecules* **8,** No. 5, 638-641 (Sept.-Oct. 1975).

Santoro, A., Choi, C. S., Abel, J. E., **1,5-diacetyl-3,7-dinitro-1,3,5,7-tetraazacyclooctane (DADN),** *Acta Crystallogr.* **B31,** Part 8, 2126-2128 (Aug. 1975).

Scheide, E. P., Taylor, J. K., **A piezoelectric crystal dosimeter for monitoring mercury vapor in industrial atmospheres,** *Am. Ind. Hyg. Assoc. J.* **36,** No. 12, 897-901 (Dec. 1975).

Sengers, J. M. H. L., Sengers, J. V., **Universality of critical behavior in gases,** *Phys. Rev. A* **12,** No. 6, 2622-2627 (Dec. 1975).

Shank, R., Henderson, M., **Federal Library cooperation,** *Library Trends* **24,** No. 2, 277-292 (Oct. 1975).

Sibley, E. H., **On the equivalences of data based systems,** *Proc. Association for Computing Machinery SIGMOD Conf., Ann Arbor, Mich., May 1-3, 1974,* **II,** 1-52 (1974).

Sibley, E. H., **Summary of CODASYL report on selection and acquisition of data base management systems,** *Digest of Papers, Association for Computing Machinery 1975 Conf., Minneapolis, Minn., Sept. 1974,* 3 pages (1974).

Simiu, E., Filliben, J. J., **Structural safety and the probabilistic definition of design wind speeds,** *Proc. CIB Int. Symp. on the Climatology of Building, Zurich, Switzerland, Sept. 25-27, 1974,* 99 pages (1974).

Stein, S. R., **Application of superconductivity to precision oscillators,** *Proc. 29th Annual Symp. on Frequency Control, Atlantic City, N.J., May 28-30, 1975,* pp. 321-327 (1975).

Straty, G. C., **Hypersonic velocities in saturated and compressed fluid methane,** *Cryogenics* **15,** No. 12, 729-731 (Dec. 1975).

Sugar, J., **Ionization energies of quadruply ionized rare earths,** *J. Opt. Soc. Amer.* **65,** No. 11, 1366-1367 (Nov. 1975).

Teague, E. C., **Surface finish measurements: An overview,** *Soc. Manuf. Eng. Tech. Pap.* **IQ75-137,** 1-21 (1975).

Utton, D. B., **The response of phase-sensitive detectors to Lorentzian and Gaussian lineshapes,** *J. Appl. Phys.* **46,** No. 12, 5268-5271 (Dec. 1975).

Vagelatos, N., Rowe, J. M., Rush, J. J., **Lattice dynamics of ND_4I in the NaCl phase (I) at 296°K,** *Phys. Rev. B* **12,** No. 10, 4522-4529 (Nov. 15, 1975).

Ventre, F. T., **Transforming environmental research into regulatory policy,** (6th Annual Meeting of the Environmental Design Research Association (EDRA), Lawrence, Kans., Apr. 23, 1975), Paper in *Responding to Social Change,* pp. 277-284 (Dowden, Hutchinson & Ross, Inc., Stroudsburg, Pa., Apr. 1975).

Walker, G. R., Minor, J. E., Marshall, R. D., **The Darwin cyclone valuable lesson in structural design,** *Civ. Eng.* **45,** No. 12, 82-86 (Dec. 1975).

Walls, F. L., DeMarchi, A., **RF spectrum of a signal after frequency multiplication; measurement and comparison with a simple calculation,** *IEEE Trans. Instrum. Meas.* **IM-24,** 210-217 (Sept. 1975).

Walls, F. L., Wainwright, A. E., **Measurement of the short-term stability of quartz crystal resonators and the implications for crystal oscillator design and applications,** *IEEE Trans. Instrum. Meas.* **IM-24,** 15-20 (Mar. 1975).

Walls, F. L., Wainwright, A. E., **Measurements of the short-term stability of quartz crystal resonators—a window on future developments in crystal oscillators,** *Proc. 6th Annual Precise Time and Time Interval (PTTI) Planning Meeting, Washington, D.C., Dec. 3-5, 1974,* pp. 143-153 (1974).

Wang, F. W., DiMarzio, E. A., **The dynamics of block-copolymer molecules in solution. The free-draining limit,** *Macromolecules* **8,** No. 3, 356-360 (May-June 1975).

Zimmerman, J. E., **Phase slip, dissipation, Bernoulli effect, parametric capacitance, and other curious features of the Josephson effect,** *IEEE Trans. Magn.* **MAG-11,** No. 2, 852-855 (Mar. 1975).

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, and chemistry. It is published in two sections, available separately:

- **Physics and Chemistry (Section A)**

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $17.00; Foreign, $21.25.

- **Mathematical Sciences (Section B)**

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $9.00; Foreign, $11.25.

DIMENSIONS/NBS (formerly Technical News Bulletin)—This monthly magazine is published to inform scientists, engineers, businessmen, industry, teachers, students, and consumers of the latest advances in science and technology, with primary emphasis on the work at NBS. The magazine highlights and reviews such issues as energy research, fire protection, building technology, metric conversion, pollution abatement, health and safety, and consumer product performance. In addition, it reports the results of Bureau programs in measurement standards and techniques, properties of matter and materials, engineering standards and services, instrumentation, and automatic data processing.

Annual subscription: Domestic, $9.45; Foreign, $11.85.

NONPERIODICALS

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a world-wide program coordinated by NBS. Program under authority of National Standard Data Act (Public Law 90-396).

NOTE: At present the principal publication outlet for these data is the Journal of Physical and Chemical Reference Data (JPCRD) published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements available from ACS, 1155 Sixteenth St. N.W., Wash. D. C. 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The purpose of the standards is to establish nationally recognized requirements for products, and to provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

Order **above** *NBS publications from: Superintendent of Documents, Government Printing Office, Washington, D.C. 20402.*

Order **following** *NBS publications—NBSIR's and FIPS from the National Technical Information Services, Springfield, Va. 22161.*

Federal Information Processing Standards Publications (FIPS PUBS)—Publications in this series collectively constitute the Federal Information Processing Standards Register. Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NBS Interagency Reports (NBSIR)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Services (Springfield, Va. 22161) in paper copy or microfiche form.

BIBLIOGRAPHIC SUBSCRIPTION SERVICES

The following current-awareness and literature-survey bibliographies are issued periodically by the Bureau:

Cryogenic Data Center Current Awareness Service. A literature survey issued biweekly. Annual subscription: Domestic, $20.00; Foreign, $25.00.

Liquified Natural Gas. A literature survey issued quarterly. Annual subscription: $20.00.

Superconducting Devices and Materials. A literature survey issued quarterly. Annual subscription: $20.00. Send subscription orders and remittances for the preceding bibliographic services to National Bureau of Standards, Cryogenic Data Center (275.02) Boulder, Colorado 80302.

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
Washington, D.C. 20234

OFFICIAL BUSINESS
Penalty for Private Use, $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE
COM-215
SPECIAL FOURTH-CLASS RATE
BOOK

JOURNAL OF RESEARCH of the National Bureau of Standaards
B. Mathematical Sciences

Vol. 80B, No. 2 Apr.-June 1976

SJRB SERIA300S ISSDUE002R 1
SERIALS PROCESSING DEPT
XEROX UNIVERSITY MICROFILMS
300 N ZEEB RD
ANN ARBOR MI 48106

Page

The dependence of inspection-system performance on levels of penalties and inspection resources. A. J. Goldman and M. H. Pearl. 189

A minimax-measure intersection problem. Philip R. Meyers. 237

Inverting sparse matrices by tree partitioning. D. R. Shier. 245

The structure of higher degree symmetry classes of tensors. Russell Merris. 259

The Kronecker power of a permutation. Russell Merris. 265

On Iohvidov's proofs of the Fischer-Frobenius theorem. R. C. Thompson. 269

Matrix algebra and eigenvalues for the bead/spring model of polymer solutions. Jeffrey T. Fong and Anton Peterlin. 273

Predictable regular continued cotangent expansions. Jeffrey Shallit. 285

Automatic computing methods for special functions. Part III. The sine, cosine, exponential integrals, and related functions. Irene A. Stegun and Ruth Zucker. 291

Definite integrals of the complete elliptic integral K. M. L. Glasser. 313

Publications of the National Bureau of Standards. 325

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Order by SD Catalog No. C 13.22/sec.B:80B (2). Price $2.25 (single copy). Subscription price: $9.00 a year; $2.25 additional for foreign mailing.
UNITED STATES GOVERNMENT PRINTING OFFICE, WASHINGTON : 1976